Analisis Segmentasi Calon Pemegang Kartu Kredit Berdasarkan Karakteristik Finansial Menggunakan Algoritma K-Means

Shabiha Rahma Fauziah / 1206220017

Link Github: https://github.com/shabiharf/Finalproject-ML

1. Pendahuluan

1.1 Latar Belakang

Dalam industri keuangan, khususnya pada sektor perbankan dan penerbitan kartu kredit, proses evaluasi kelayakan calon nasabah merupakan aspek yang sangat penting. Setiap individu yang mengajukan kartu kredit memiliki latar belakang dan kondisi finansial yang beragam. Oleh karena itu, diperlukan pemahaman yang lebih mendalam terhadap karakteristik setiap calon pengguna agar proses persetujuan kartu kredit dapat dilakukan secara akurat dan adil.

Seiring dengan meningkatnya jumlah pengajuan dan volume data historis calon pengguna, pendekatan manual dalam proses evaluasi menjadi tidak efisien dan berisiko tinggi terhadap kesalahan atau bias. Dalam situasi seperti ini, diperlukan solusi berbasis teknologi yang mampu mengelompokkan data pengguna secara otomatis berdasarkan kemiripan karakteristik. Salah satu pendekatan yang dapat digunakan adalah metode clustering dalam pembelajaran mesin (machine learning).

Clustering merupakan teknik unsupervised learning yang digunakan untuk mengelompokkan data ke dalam beberapa kelompok berdasarkan kemiripan tertentu antar data. Dalam konteks ini, algoritma K-Means menjadi salah satu metode clustering yang paling umum digunakan karena kemampuannya dalam menemukan pola dan segmentasi tersembunyi pada data dengan efisien.

Dengan menerapkan K-Means Clustering pada data historis calon pemegang kartu kredit, perusahaan dapat memperoleh segmentasi pengguna yang lebih terstruktur, mulai dari pengguna dengan risiko rendah hingga pengguna dengan potensi gagal bayar yang tinggi. Informasi ini dapat digunakan untuk mendukung keputusan persetujuan kredit, menentukan batasan limit awal, serta menyusun strategi pemasaran dan manajemen risiko yang lebih tepat sasaran.

Melalui proyek ini, dilakukan proses analisis segmentasi terhadap calon pengguna kartu kredit menggunakan metode K-Means Clustering. Tahapan yang dilakukan mencakup eksplorasi data, evaluasi fitur, normalisasi, pemilihan jumlah klaster optimal, visualisasi hasil clustering, hingga interpretasi masing-masing segmen. Diharapkan hasil akhir dari analisis ini dapat memberikan insight yang bermanfaat bagi perusahaan dalam menyusun kebijakan berbasis data (data-driven decision making).

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka rumusan masalah yang ingin dijawab dalam proyek ini adalah sebagai berikut:

1. Berapa rata-rata usia, penghasilan, dan lama bekerja calon pemegang kartu kredit?
2. Apakah calon pemegang kartu kredit yang disetujui memiliki ciri khas tertentu dari sisi penghasilan, skor kredit, dan pengalaman kerja?
3. Apakah terdapat hubungan yang kuat antar fitur numerik yang digunakan dalam penilaian kelayakan pengguna?

4. Bagaimana cara mengelompokkan calon pemegang kartu kredit menjadi beberapa segmen berdasarkan data mereka, dan seperti apa ciri-ciri tiap kelompoknya?

1.3 Tujuan Proyek:

Adapun tujuan dari proyek analisis ini adalah:

1. Mengetahui statistik dasar seperti usia, penghasilan, dan pengalaman kerja rata-rata calon pemegang kartu kredit.
2. Menganalisis perbedaan profil antara pengguna yang disetujui dan tidak disetujui dari sisi fitur-fitur penting.
3. Mengidentifikasi sejauh mana hubungan antar fitur numerik saling memengaruhi.
4. Melakukan clustering terhadap calon pengguna kartu kredit untuk mengidentifikasi segmen potensial berdasarkan karakteristik mereka secara menyeluruh.

2. Eksplorasi Data (EDA)

2.1 Deskripsi Dataset

Dataset yang digunakan dalam proyek ini merupakan data historis calon pemegang kartu kredit yang telah melalui proses pembersihan (cleaned dataset). Dataset ini terdiri dari 690 baris (data calon pengguna) dan 12 kolom (fitur), di mana setiap baris mewakili satu individu calon pengguna kartu kredit. Setiap fitur mencerminkan aspek tertentu dari karakteristik demografis dan kondisi finansial calon pengguna. Fitur-fitur ini digunakan sebagai dasar untuk mengevaluasi kelayakan dalam proses pengajuan kartu kredit.

Berikut adalah deskripsi singkat dari fitur-fitur utama yang digunakan:

- Age: Usia calon pengguna kartu kredit (dalam tahun).
- Income: Pendapatan per tahun dalam satuan ribuan dolar.
- YearsEmployed: Lama bekerja dalam satuan tahun.
- Debt: Jumlah total utang saat ini.
- CreditScore: Skor kredit yang mencerminkan reputasi finansial.
- PriorDefault: Status apakah pernah mengalami gagal bayar sebelumnya (0 = Tidak, 1 = Ya).
- Employed: Status apakah saat ini sedang bekerja (0 = Tidak, 1 = Ya).
- Married: Status pernikahan (0 = Tidak, 1 = Ya).
- BankCustomer: Apakah merupakan nasabah bank (0 = Tidak, 1 = Ya).
- ZipCode: Kode pos tempat tinggal.
- Approved: Status pengajuan kartu kredit (0 = Ditolak, 1 = Disetujui).
- Cluster_Relabeled: Label hasil clustering (0–3) berdasarkan karakteristik finansial.

Sumber asli data berasal dari UCI Machine Learning Repository, dengan judul Credit Approval Dataset, yang dapat diakses melalui tautan: https://archive.ics.uci.edu/dataset/27/credit+approval.

Dataset yang digunakan dalam proyek ini merupakan versi yang telah dibersihkan dan dipublikasikan ulang oleh pengguna Kaggle melalui tautan: https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data/data. Oleh karena itu, data yang digunakan sudah bebas dari nilai kosong dan telah dikonversi sepenuhnya ke format numerik sehingga dapat langsung digunakan untuk eksplorasi dan analisis.

2.2 Prapemrosesan Data

Sebelum dilakukan proses analisis dan pengelompokan, data perlu melalui beberapa tahap pra-pemrosesan untuk memastikan bahwa data dalam kondisi bersih, relevan, dan siap digunakan dalam pemodelan machine learning. Proses ini mencakup seleksi fitur, penanganan nilai kosong, transformasi nilai kategorik, dan normalisasi. Langkah-langkah pra-pemrosesan yang dilakukan adalah sebagai berikut:

1. **Seleksi Fitur**

   Tidak semua fitur dalam dataset digunakan secara langsung. Hanya fitur-fitur yang relevan dan dapat memberikan informasi penting terhadap karakteristik calon pengguna kartu kredit yang dipertahankan. Untuk memperkuat seleksi, dilakukan uji statistik menggunakan independent t-test guna mengukur signifikansi perbedaan antar fitur terhadap status persetujuan.

2. **Pemeriksaan dan Penanganan Nilai Kosong**

   Dataset yang digunakan merupakan versi yang telah dibersihkan, sehingga tidak ditemukan nilai kosong (missing values) dalam fitur-fitur yang tersedia. Pemeriksaan ini dilakukan untuk memastikan integritas data.

3. **Transformasi Fitur Kategorik**

   Beberapa fitur dalam dataset berupa data kategorik biner (seperti PriorDefault, Married, Employed, BankCustomer). Seluruh fitur kategorik tersebut telah dikonversi menjadi bentuk numerik (0 atau 1) agar dapat diproses lebih lanjut oleh algoritma machine learning.

4. **Normalisasi Skala Fitur**

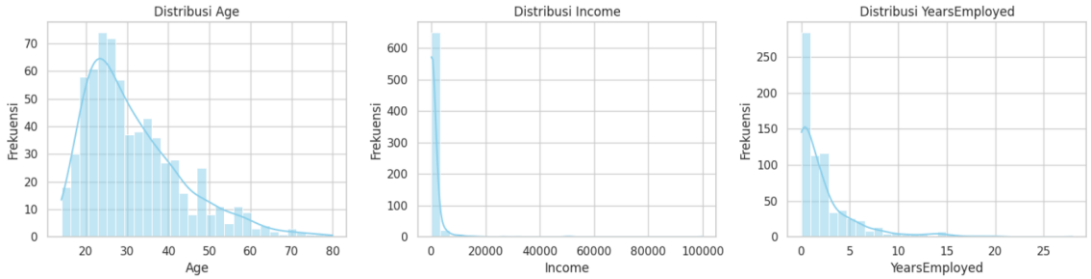
   Mengingat adanya perbedaan skala antar fitur numerik, seperti Income, Age, dan CreditScore, maka dilakukan proses normalisasi menggunakan metode Min-Max Scaling secara manual. Hal ini bertujuan agar semua fitur memiliki rentang nilai yang seragam (0 hingga 1), sehingga tidak ada fitur yang mendominasi perhitungan jarak dalam proses clustering.

5. **Penyimpanan Dataset Final**

   Setelah semua proses pra-pemrosesan dilakukan, dataset akhir yang telah dinormalisasi dan diseleksi fitur-fiturnya kemudian digunakan sebagai input dalam proses clustering menggunakan algoritma K-Means.

2.3 Visualisasi dan Temuan Awal

a. Rata-rata usia, penghasilan, dan lama bekerja calon pemegang kartu kredit



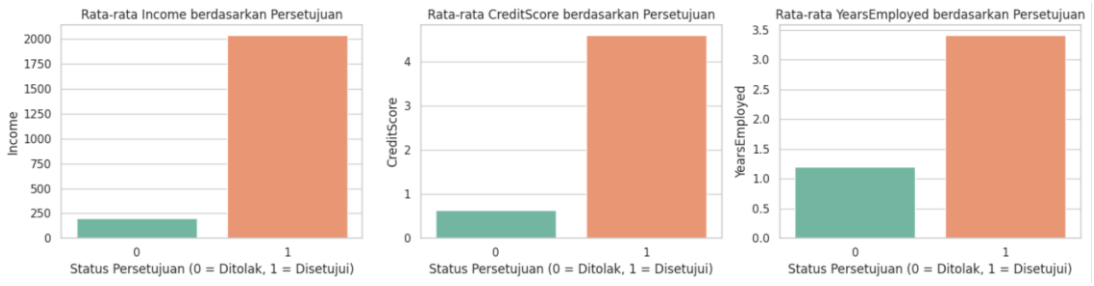
Fitur	Rata-Rata
Age	31.51
Income	1017.39
YearsEmployed	2.18

Berdasarkan hasil analisis statistik deskriptif terhadap fitur-fitur numerik, diperoleh bahwa rata-rata usia calon pemegang kartu kredit adalah sekitar 31,5 tahun, dengan penghasilan rata-rata sebesar USD 1017.39 serta rata-rata lama bekerja selama 2,18 tahun. Nilai-nilai ini memberikan gambaran awal mengenai profil umum populasi dalam dataset.

Melalui visualisasi distribusi menggunakan histogram dan kurva kepadatan (KDE), ditemukan bahwa fitur Age memiliki sebaran yang relatif simetris. Hal ini terlihat dari nilai rata-rata dan median yang hampir sama, menunjukkan bahwa kelompok usia dalam data tersebar merata di sekitar nilai tengah. Sebaliknya, dua fitur lainnya yaitu Income dan YearsEmployed menunjukkan pola distribusi yang condong ke kanan (right-skewed). Artinya, mayoritas calon pengguna kartu kredit memiliki pendapatan dan pengalaman kerja yang tergolong rendah, namun terdapat sejumlah kecil individu dengan pendapatan dan masa kerja yang sangat tinggi. Kondisi ini menyebabkan nilai rata-rata menjadi lebih besar dibandingkan median.

Pola distribusi yang tidak simetris ini penting untuk dicermati karena dapat memengaruhi proses analisis lebih lanjut, terutama pada tahap normalisasi data. Nilai-nilai ekstrem (outlier) pada pendapatan dan masa kerja dapat mempengaruhi hasil clustering apabila tidak ditangani dengan baik. Oleh karena itu, informasi statistik dasar dan pola distribusi ini menjadi landasan penting dalam memahami struktur awal data dan mempersiapkan data untuk proses segmentasi yang lebih akurat.

b. Calon pemegang kartu kredit yang disetujui memiliki ciri khas tertentu dari sisi penghasilan, skor kredit, dan pengalaman kerja



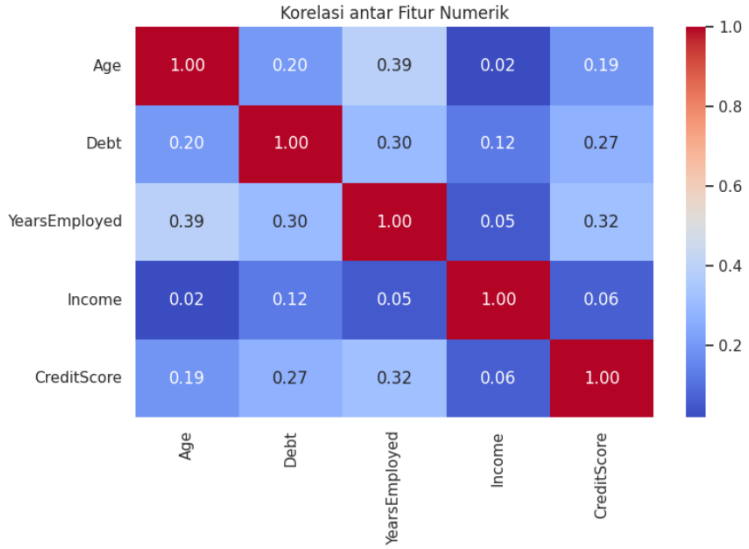
Berdasarkan hasil visualisasi komparatif antara pengguna yang disetujui dan yang tidak disetujui pengajuan kartu kreditnya, ditemukan adanya perbedaan karakteristik yang cukup signifikan. Analisis ini dilakukan terhadap tiga fitur utama, yaitu Income, CreditScore, dan YearsEmployed.

Rata-rata penghasilan calon pengguna yang disetujui terlihat lebih tinggi dibandingkan dengan mereka yang ditolak. Hal serupa juga berlaku pada fitur CreditScore dan YearsEmployed, di mana pengguna yang disetujui umumnya memiliki skor kredit yang lebih baik serta pengalaman kerja yang lebih lama. Pola ini menunjukkan adanya kecenderungan bahwa individu dengan kondisi finansial yang lebih stabil dan riwayat kredit yang baik lebih berpeluang untuk disetujui dalam proses pengajuan kartu kredit.

Perbedaan ini menjadi bukti awal bahwa fitur-fitur seperti penghasilan, skor kredit, dan pengalaman kerja berperan penting dalam proses penilaian kelayakan

calon pengguna kartu kredit. Informasi ini juga dapat dimanfaatkan pada tahap segmentasi dan penyusunan strategi risiko ke depan.

c. Korelasi antar fitur numerik yang digunakan dalam penilaian kelayakan pengguna

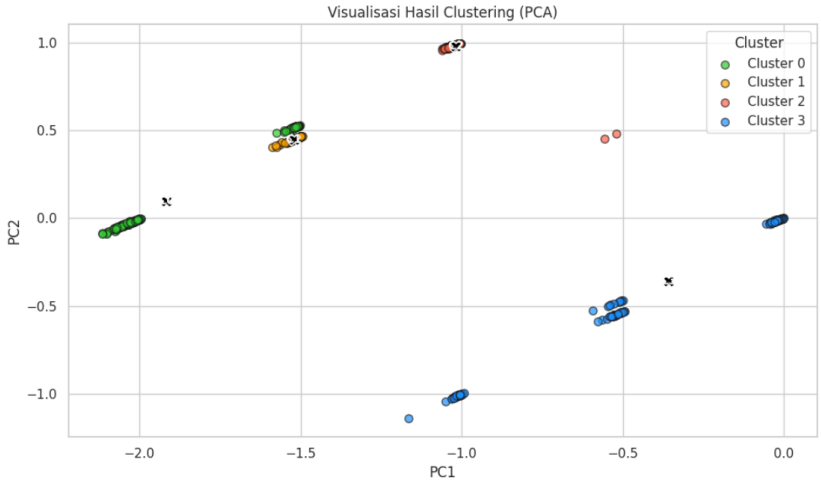


Untuk mengetahui apakah terdapat hubungan antar fitur numerik dalam dataset, dilakukan analisis korelasi. Korelasi ini menunjukkan seberapa erat hubungan antara dua variabel numerik. Nilainya berkisar antara -1 hingga 1, di mana nilai mendekati 1 berarti kedua variabel bergerak searah dan sangat berkaitan, sementara nilai mendekati 0 menunjukkan bahwa tidak ada hubungan yang berarti.

Hasil analisis menunjukkan bahwa sebagian besar fitur numerik dalam dataset memiliki hubungan yang lemah satu sama lain. Sebagai contoh, korelasi antara fitur Income dan YearsEmployed hanya sebesar 0.05. Artinya, dalam data ini, penghasilan seseorang tidak selalu meningkat seiring dengan lama bekerja, mungkin karena adanya pengaruh faktor lain seperti jenis pekerjaan, tingkat pendidikan, atau sektor industri. Sementara itu, CreditScore dan YearsEmployed memiliki korelasi sebesar 0.32, yang berarti semakin lama seseorang bekerja, biasanya skor kreditnya cenderung sedikit lebih baik, meskipun hubungan ini tidak terlalu kuat. Fitur Debt juga menunjukkan korelasi yang lemah dengan fitur lain, termasuk dengan Income dan CreditScore, yang mengindikasikan bahwa tingkat utang seseorang tidak selalu sejalan dengan pendapatan atau reputasi keuangannya.

Secara keseluruhan, tidak ditemukan korelasi yang sangat kuat antar fitur-fitur numerik yang ada. Hal ini merupakan kondisi yang baik untuk analisis clustering, karena masing-masing fitur dapat memberikan informasi unik yang tidak tumpang tindih satu sama lain. Dengan demikian, proses pengelompokan akan lebih kaya informasi dan tidak didominasi oleh satu fitur tertentu saja.

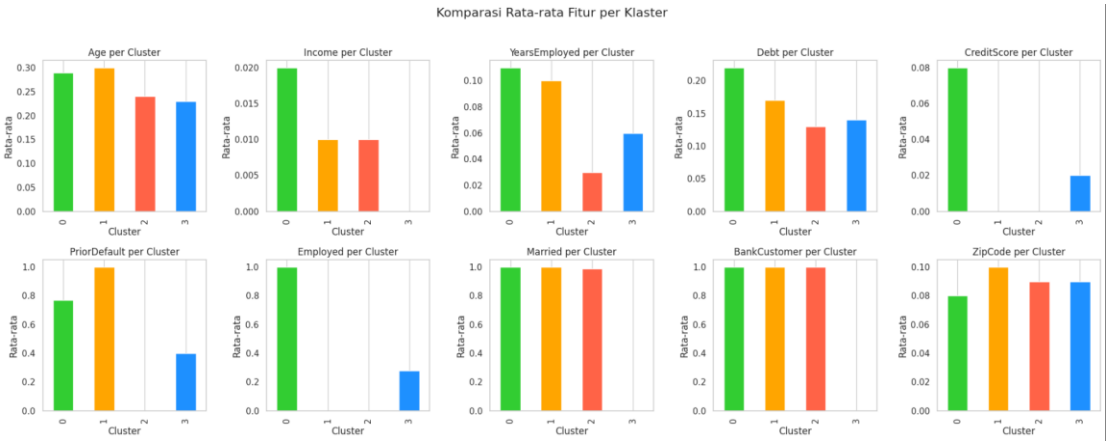
d. Kelompok calon pemegang kartu kredit



Visualisasi hasil clustering yang dilakukan menggunakan PCA (Principal Component Analysis) memberikan wawasan yang mendalam dengan memetakan data berdimensi tinggi ke dalam dua komponen utama, yaitu PC1 dan PC2. Proses ini memungkinkan kita untuk mengamati pola klaster antar calon pengguna kartu kredit secara visual dan intuitif. Dari grafik yang dihasilkan, terlihat bahwa data berhasil dikelompokkan ke dalam empat klaster yang cukup terpisah secara visual.

Masing-masing klaster memiliki area dominan tersendiri di ruang dua dimensi, yang mengindikasikan bahwa fitur-fitur penting seperti pendapatan, skor kredit, histori default, dan status pekerjaan berkontribusi secara signifikan dalam pembentukan klaster. Titik centroid yang ditandai dengan simbol "X" menunjukkan pusat dari masing-masing klaster, dan posisi centroid yang berjauhan memperkuat argumen bahwa klaster tersegmentasi dengan baik. Penggunaan warna yang berbeda untuk setiap klaster, seperti limegreen, orange, tomato, dan dodgerblue, tidak hanya mempercantik visualisasi tetapi juga membantu dalam mengidentifikasi keberagaman profil pengguna. Terlihat bahwa tidak banyak data point yang tumpang tindih antar klaster, yang menandakan bahwa struktur klaster cukup jelas dan kuat. Dengan demikian, visualisasi ini tidak hanya memberikan gambaran yang jelas tentang pengelompokan data, tetapi juga menyoroti faktor-faktor kunci yang mempengaruhi perilaku dan karakteristik calon pengguna kartu kredit, yang dapat menjadi dasar yang kuat untuk pengambilan keputusan strategis dalam pemasaran dan pengembangan produk kartu kredit di masa depan.

e. Karakteristik kelompok calon pemegang kartu kredit



Berdasarkan visualisasi komparatif dan perhitungan rata-rata nilai fitur untuk masing-masing klaster yang telah dinormalisasi, kita dapat mengidentifikasi karakteristik unik dari tiap segmen calon pemegang kartu kredit.

Klaster 0 – "Segmen Elite Finansial" menunjukkan pendapatan dan lama bekerja yang sangat tinggi, dengan skor kredit tertinggi di antara semua klaster. Mereka tidak memiliki histori gagal bayar (PriorDefault = 0) dan mayoritas dari mereka sudah bekerja dan menikah. Klaster ini sangat cocok dijadikan target prioritas untuk persetujuan dan penawaran produk premium.

Klaster 1 – "Segmen Stabil" memiliki pendapatan dan lama kerja di atas rata-rata, serta skor kredit yang cukup baik dengan frekuensi gagal bayar yang jarang. Mayoritas dari mereka memiliki pekerjaan dan pernah menikah, menjadikan calon pengguna ini aman dan layak untuk diberikan limit kredit sedang hingga tinggi.

Klaster 2 – "Segmen Rentan" ditandai dengan pendapatan dan skor kredit yang rendah. Rata-rata, mereka memiliki histori gagal bayar (PriorDefault mendekati 1) dan tingkat utang yang cukup tinggi. Banyak dari mereka yang belum bekerja dan belum menikah, sehingga klaster ini perlu diseleksi dengan ketat dan mungkin memerlukan edukasi finansial untuk meningkatkan pemahaman mereka tentang manajemen keuangan.

Klaster 3 – "Segmen Muda atau Pasif" terdiri dari individu dengan usia dan pendapatan yang rendah. Meskipun mereka memiliki utang yang rendah, skor kredit mereka juga rendah, kemungkinan disebabkan oleh kurangnya aktivitas kredit. Lama bekerja yang pendek menunjukkan bahwa mereka bisa jadi merupakan pengguna baru atau belum memiliki histori kredit yang signifikan. Klaster ini memiliki potensi untuk tumbuh dalam jangka panjang, tetapi saat ini belum cocok untuk diberikan limit kredit yang besar.

Dengan pemahaman yang mendalam tentang karakteristik masing-masing klaster, lembaga keuangan dapat merancang strategi pemasaran dan penawaran produk yang lebih tepat sasaran, serta meningkatkan pengalaman pengguna dalam pengelolaan keuangan mereka.

3. Metode Penjelasan

3.1 Pra-pemrosesan Teks

Tahapan pra-pemrosesan data merupakan langkah penting dalam analisis sentimen berbasis teks. Tujuannya adalah untuk membersihkan dan menyiapkan data teks agar

dapat digunakan secara efektif oleh model machine learning. Pada proyek ini, tahapan pra-pemrosesan yang dilakukan meliputi:

1. Pemilihan Fitur Numerik

   Pada tahap awal, dilakukan pemilihan fitur numerik dan kategorik yang signifikan berdasarkan hasil uji statistik independen t-test. Pemilihan ini bertujuan untuk memastikan bahwa hanya fitur-fitur yang memiliki pengaruh signifikan terhadap variabel target yang akan dipertimbangkan dalam analisis lebih lanjut, sehingga mengoptimalkan kualitas data yang digunakan.
2. Pengecekan Data Kosong (Missing Value)

   Setelah fitur dipilih, dataset diperiksa untuk mendeteksi adanya nilai yang hilang (missing values). Identifikasi ini sangat penting karena keberadaan data kosong dapat mengganggu hasil analisis.
3. Normalisasi Skala

   Agar semua fitur memiliki skala yang seragam, dilakukan normalisasi dengan metode Min-Max secara manual. Hal ini sangat penting terutama untuk algoritma clustering seperti K-Means yang sensitif terhadap perbedaan skala antar variabel. Dengan normalisasi, semua fitur diubah ke rentang nilai yang sama, sehingga mencegah dominasi fitur tertentu dalam proses pembentukan klaster.
4. Analisisi Korelasi Antar Fitur

   Selanjutnya, analisis korelasi antar fitur dilakukan untuk memahami hubungan dan ketergantungan antar variabel. Tujuannya adalah memastikan bahwa fitur yang digunakan tidak memiliki redundansi yang tinggi dan masing-masing memberikan informasi unik yang dapat memperkaya model clustering.

3.2 Algoritma Klustering

a. K-Means Clustering

Algoritma K-Means Clustering adalah metode unsupervised learning yang populer untuk mengelompokkan data ke dalam beberapa klaster berdasarkan kemiripan fitur-fitur yang dimiliki oleh data tersebut. Prinsip dasar dari K-Means adalah membagi dataset menjadi K klaster, di mana setiap data dimasukkan ke dalam klaster dengan jarak terdekat ke pusat klaster, yang disebut centroid. Secara matematis, jarak yang biasanya digunakan adalah jarak Euclidean, yang dihitung dengan rumus:

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - c_{jm})^2}$$

di mana x_i adalah titik data ke-i dengan M fitur, dan Cj adalah centroid klaster ke-j. Algoritma ini bekerja dengan cara menginisialisasi centroid secara acak, kemudian melakukan iterasi berulang.

Pada setiap iterasi, setiap data akan diberi label klaster berdasarkan jarak minimum ke centroid, setelah itu centroid diperbarui dengan menghitung rata-rata dari seluruh titik data yang tergabung dalam klaster tersebut:

$$c_j = \frac{1}{n_j} \sum_{i \in C_j} x_i$$

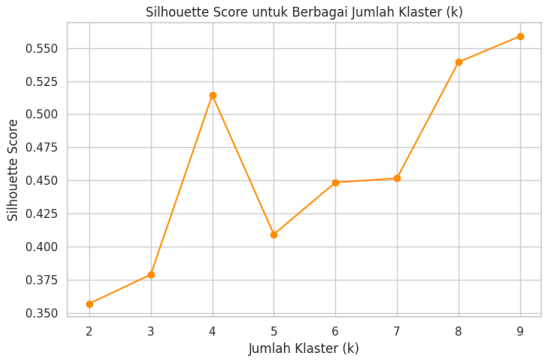
adalah jumlah data pada klaster c_j.

Dalam proyek "Analisis Segmentasi Calon Pemegang Kartu Kredit Berdasarkan Karakteristik Finansial", kami menggunakan fitur-fitur seperti 'Age', 'Income', 'YearsEmployed', 'Debt', 'CreditScore', 'PriorDefault', 'Employed', 'Married', 'BankCustomer', dan 'ZipCode' untuk mengelompokkan data calon pemegang kartu kredit. Dengan menentukan jumlah klaster sebanyak empat berdasarkan evaluasi metode Elbow dan Silhouette Score, algoritma K-Means secara iteratif mengelompokkan data berdasarkan kedekatan fitur tersebut ke centroid masing-masing klaster. Sebagai contoh sederhana, jika terdapat satu data calon pemegang kartu kredit dengan nilai fitur tertentu, maka jarak Euclidean antara data tersebut dengan setiap centroid dihitung. Data akan dimasukkan ke klaster dengan jarak terkecil ke centroid tersebut. Proses pembaruan centroid dan pengelompokan ulang data ini diulang hingga tidak terjadi perubahan signifikan pada posisi centroid, sehingga diperoleh segmentasi yang stabil dan meaningful berdasarkan karakteristik finansial calon pengguna kartu kredit.

4. Hasil dan Pembahasan

Dalam upaya menentukan jumlah klaster optimal untuk segmentasi calon pemegang kartu kredit, telah dilakukan evaluasi komprehensif menggunakan kombinasi dua metode populer: Elbow Method dan Silhouette Score.

4.1 Visualisasi Evaluasi Algoritma: "Silhouette Score untuk Berbagai Jumlah Klaster (k)

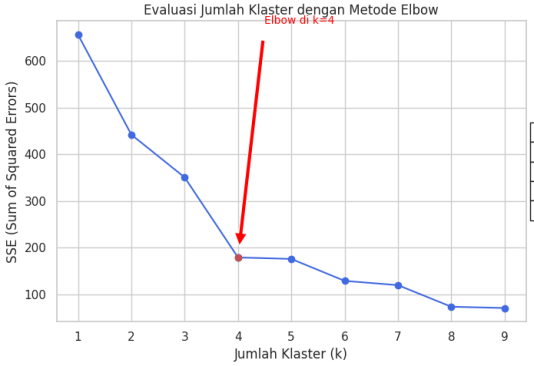


Jumlah Klaster (k)	Silhouette Score
2.0	0.3573
3.0	0.3791
4.0	0.5145
5.0	0.4095
6.0	0.4486
7.0	0.4517
8.0	0.5393
9.0	0.5588

Visualisasi ini, yang disajikan dalam bentuk grafik garis dan tabel, menunjukkan performa algoritma K-Means berdasarkan nilai Silhouette Score untuk berbagai jumlah klaster (k) yang berbeda. Silhouette Score adalah metrik penting yang mengukur seberapa baik setiap data point cocok dengan klasternya sendiri dibandingkan dengan

klaster tetangganya, dengan rentang nilai dari -1 hingga 1. Nilai yang lebih tinggi mengindikasikan klaster yang lebih padat dan terpisah dengan baik. Dari grafik, terlihat bahwa nilai Silhouette Score berfluktuasi seiring dengan peningkatan jumlah klaster. Penting untuk dicatat bahwa nilai tertinggi tercapai pada k=9 dengan skor 0.5588, namun terdapat juga puncak signifikan pada k=4 dengan skor 0.5145 dan pada k=8 dengan skor 0.5393. Analisis ini memberikan gambaran awal mengenai potensi jumlah klaster yang optimal berdasarkan kualitas pemisahan klaster.

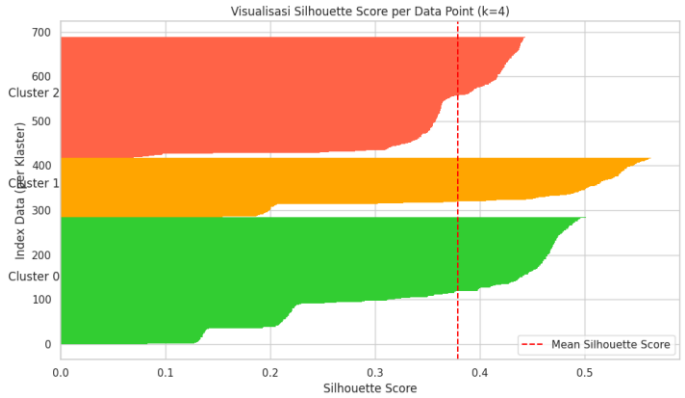
4.2 Evaluasi Jumlah Klaster dengan Metode Elbow



Jumlah Klaster (k)	SSE	Silhouette Score
2.0	441.93	0.3573
4.0	179.17	0.5145
6.0	129.03	0.4486
8.0	73.68	0.5393

Visualisasi ini, yang dikenal sebagai Elbow Curve, menyajikan nilai Sum of Squared Errors (SSE) atau inersia untuk setiap jumlah klaster (k). SSE mengukur total jarak kuadrat antara setiap titik data dan centroid klaster-nya, sehingga nilai SSE yang lebih rendah menunjukkan klaster yang lebih padat. Tujuan dari metode ini adalah mengidentifikasi "titik siku" (elbow point) pada grafik, di mana penurunan SSE mulai melandai secara signifikan. Dari grafik, terlihat bahwa penurunan SSE sangat drastis dari k=1 hingga k=4. Setelah k=4, penurunan SSE menjadi tidak terlalu curam. Panah merah pada grafik secara jelas menunjuk pada k=4 sebagai titik siku, mengindikasikan bahwa menambah jumlah klaster setelah k=4 tidak lagi memberikan pengurangan error yang signifikan. Ini menunjukkan k=4 sebagai kandidat kuat untuk jumlah klaster yang efisien.

4.3 Visualisasi Silhouette Score per Data Point (k=4)



Untuk memperkuat pemilihan k=4 sebagai jumlah klaster optimal, visualisasi Silhouette Score per data point ini menjadi sangat penting. Grafik ini menampilkan nilai Silhouette Score individu untuk setiap data point dalam masing-

masing klaster ketika jumlah klaster diatur menjadi 4. Setiap batang horizontal merepresentasikan satu data point, dan panjangnya menunjukkan nilai Silhouette Score-nya. Batang yang lebih panjang ke kanan mengindikasikan bahwa data point tersebut sangat cocok berada dalam klasternya. Garis putus-putus merah vertikal menunjukkan rata-rata Silhouette Score keseluruhan untuk k=4. Dari visualisasi ini, terlihat bahwa sebagian besar data point di setiap klaster memiliki Silhouette Score positif dan relatif tinggi, yang menunjukkan bahwa data-data tersebut tergabung dengan baik dalam klaster mereka dan terpisah dengan jelas dari klaster lain. Ini mengkonfirmasi bahwa struktur klaster saat k=4 tidak hanya efisien berdasarkan Elbow Method, tetapi juga berkualitas tinggi dalam hal kohesi intra-klaster dan separasi antar-klaster, serta relatif stabil tanpa banyak tumpang tindih antar klaster yang signifikan.

4.4 Evaluasi Jumlah Klaster: Kombinasi Elbow Method dan Silhouette Score

Untuk menentukan jumlah klaster yang optimal, dilakukan evaluasi menggunakan dua pendekatan: Elbow Method yang mengukur total error (SSE) untuk melihat seberapa besar penyebaran data dalam klaster, dan Silhouette Score yang mengukur seberapa baik data dikelompokkan (rapat dalam klaster, jauh dari klaster lain).

Dari grafik Elbow Curve, terlihat bahwa penurunan SSE sangat signifikan dari k=1 hingga k=4. Setelah k=4, penurunan mulai melandai, yang menunjukkan bahwa k=4 adalah "titik siku" (elbow point) — yaitu titik di mana menambah jumlah klaster tidak lagi menghasilkan pengurangan error yang signifikan.

Meskipun tabel evaluasi Silhouette Score menunjukkan nilai tertinggi pada k=9 dan juga k=8, nilai Silhouette Score pada k=4 juga cukup tinggi, yaitu 0.5145. Ketika dikombinasikan dengan hasil dari Elbow Method yang menunjuk pada k=4 sebagai titik efisiensi, ini menunjukkan bahwa klastering dengan k=4 tidak hanya efisien dari sisi SSE, tetapi juga memiliki kualitas klaster yang baik dari sisi pemisahan antar klaster. Meskipun pada k=6 atau k=8 skor Silhouette mendekati atau sedikit lebih tinggi, kompleksitas model meningkat dan interpretasi klaster menjadi lebih sulit. Oleh karena itu, secara keseluruhan, k=4 adalah pilihan paling seimbang antara efisiensi (SSE minimal), kualitas segmentasi (Silhouette tinggi), serta kemudahan interpretasi dan implementasi bisnis.

Pemilihan 4 klaster dianggap sebagai konfigurasi terbaik untuk mengelompokkan calon pengguna kartu kredit dalam studi ini. Visualisasi Silhouette Score per data point untuk k=4 lebih lanjut menunjukkan nilai silhouette untuk setiap data point dalam masing-masing klaster. Nilai yang lebih tinggi menandakan bahwa data point tersebut cocok berada dalam klasternya, dengan batang yang lebih panjang ke kanan menunjukkan kecocokan yang semakin baik. Garis merah putus-putus mewakili rata-rata Silhouette Score keseluruhan. Dari grafik ini, terlihat bahwa sebagian besar data memiliki skor positif dan cukup tinggi, menandakan bahwa struktur klaster saat k=4 sudah cukup baik dan stabil. Oleh karena itu, visualisasi Silhouette digunakan untuk mengkonfirmasi bahwa k=4 bukan hanya efisien (dari Elbow), tetapi juga stabil dan berkualitas tinggi.

5. Kesimpulan

Proyek analisis segmentasi ini telah berhasil mengelompokkan calon pemegang kartu kredit ke dalam **empat segmen utama** yang berbeda, didasarkan pada data historis dan beragam karakteristik finansial mereka. Keberhasilan ini dicapai melalui penerapan algoritma K-Means, sebuah metode *unsupervised learning* yang kuat, dipadukan dengan proses evaluasi yang menyeluruh. Penentuan jumlah klaster optimal sebanyak empat didasarkan pada konvergensi hasil dari dua teknik evaluasi kunci: **Elbow Method**, yang secara efektif mengidentifikasi titik efisiensi di mana penambahan klaster tidak lagi menghasilkan pengurangan *error* yang signifikan, dan **Silhouette Score**, yang mengonfirmasi kualitas pemisahan dan kekompakan klaster secara internal.

Setiap klaster yang terbentuk menunjukkan **profil karakteristik yang unik dan berbeda secara signifikan**, memberikan wawasan mendalam mengenai keragaman populasi calon pemegang kartu kredit. Segmentasi ini berhasil mengidentifikasi spektrum calon nasabah, mulai dari **segmen "elite" yang memiliki kelayakan kredit sangat tinggi** dan ideal untuk penawaran produk premium dengan limit besar, hingga **segmen "rentan" yang menunjukkan indikator risiko gagal bayar yang tinggi** dan memerlukan pendekatan yang lebih hati-hati atau penawaran produk yang disesuaikan. Visualisasi hasil *clustering* menggunakan Principal Component Analysis (PCA) secara efektif menggambarkan pemisahan yang jelas antar klaster dalam ruang berdimensi rendah, sementara analisis rata-rata fitur per klaster lebih lanjut menegaskan perbedaan profil finansial dan perilaku di antara setiap segmen. Hal ini menunjukkan bahwa **struktur klaster yang dihasilkan tidak hanya jelas secara statistik tetapi juga sangat relevan dan dapat diinterpretasikan secara bisnis.**

Secara keseluruhan, hasil segmentasi ini membuka **peluang strategis yang substansial** bagi institusi keuangan. Dengan memahami keragaman calon pengguna kartu kredit secara lebih mendalam melalui segmen-segmen yang terdefinisi ini, perusahaan dapat menyusun **strategi pemasaran dan pengelolaan risiko yang jauh lebih personal, akurat, dan terukur.** Ini akan memungkinkan alokasi sumber daya yang lebih efisien, pengembangan produk yang lebih tepat sasaran, serta mitigasi risiko yang lebih proaktif, pada akhirnya berkontribusi pada peningkatan profitabilitas dan pertumbuhan bisnis yang berkelanjutan.

6. Rekomendasi Strategis

Berdasarkan hasil analisis *clustering* yang telah mengidentifikasi empat segmen calon pemegang kartu kredit yang berbeda, berikut adalah rekomendasi strategis yang dapat diterapkan oleh perusahaan kartu kredit untuk mengoptimalkan operasional dan meningkatkan kinerja bisnis:

1. Strategi Pemasaran dan Produk Berbasis Segmen Klaster:

Target Prioritas (Klaster 0 & 1):

- **Penawaran Produk Premium:** Kelompok ini menunjukkan karakteristik finansial yang sangat kuat dan potensi loyalitas tinggi. Perusahaan direkomendasikan untuk secara proaktif mengajukan penawaran kartu kredit dengan limit menengah hingga tinggi, bahkan produk premium eksklusif.
- **Program Loyalitas dan Insentif:** Tingkatkan daya tarik dengan menawarkan program loyalitas yang menguntungkan, *cashback* yang kompetitif, diskon eksklusif, atau benefit perjalanan/gaya hidup yang sesuai dengan profil mereka.

- **Saluran Komunikasi Eksklusif:** Gunakan saluran komunikasi premium dan personal, seperti email pribadi dari manajer hubungan, telemarketing eksklusif, atau undangan ke acara khusus, untuk membangun hubungan yang lebih kuat dan meningkatkan *conversion rate*.
- **Fokus pada *Customer Lifetime Value* (CLTV):** Pertimbangkan segmen ini sebagai aset jangka panjang dan investasi dalam membangun loyalitas yang kuat untuk memaksimalkan nilai seumur hidup nasabah.

Klaster Rentan (Klaster 2):

- **Penilaian Risiko Diperketat:** Untuk segmen ini, proses penilaian risiko harus lebih ketat dan komprehensif sebelum menyetujui pengajuan kartu kredit. Pertimbangkan penggunaan model *credit scoring* yang lebih granular.
- **Limit Rendah dengan Pengawasan Ketat:** Jika pengajuan disetujui, berikan limit kartu kredit yang relatif rendah dan terapkan monitoring transaksi serta riwayat pembayaran secara berkala dan ketat untuk mendeteksi potensi risiko dini.
- **Edukasi Literasi Keuangan:** Tawarkan program edukasi literasi keuangan yang berfokus pada manajemen utang yang sehat, perencanaan anggaran, atau bahkan program pemulihan kredit bagi mereka yang menunjukkan tanda-tanda kesulitan finansial. Pendekatan ini dapat membantu mengurangi risiko *default* di masa depan.
- **Alternatif Produk Non-Kredit:** Pertimbangkan untuk menawarkan produk keuangan alternatif yang lebih aman seperti kartu debit dengan fitur khusus atau fasilitas pinjaman mikro dengan agunan, sebagai langkah awal membangun kepercayaan dan riwayat kredit positif.

Klaster Pemula/Pasif (Klaster 3):

- **Penawaran Produk *Entry-Level*:** Fokuskan strategi pada penawaran produk *entry-level* yang memiliki risiko rendah, seperti kartu debit dengan fitur bonus, kartu prabayar dengan benefit *cashback*, atau kartu kredit pemula dengan limit sangat terbatas dan persyaratan yang lebih fleksibel.
- ***Nurturing* dan Edukasi Awal:** Kembangkan program *nurturing* yang bertujuan untuk secara bertahap mendidik mereka tentang penggunaan produk keuangan yang bertanggung jawab dan membangun riwayat kredit positif. Ini dapat mencakup konten edukasi tentang manfaat penggunaan kartu kredit secara bijak dan tips membangun skor kredit.
- **Fokus pada Akuisisi dan Aktivasi Awal:** Strategi utama adalah mengakuisisi mereka dan mendorong aktivasi serta penggunaan awal untuk membangun kebiasaan transaksi, dengan harapan mereka dapat berkembang menjadi pengguna aktif yang lebih stabil di masa mendatang.

2. Optimalisasi Operasional dan Integrasi Sistem:

- **Otomatisasi Segmentasi Pengguna Baru:** Manfaatkan hasil *clustering* ini untuk mengembangkan sistem otomatisasi yang dapat mengklasifikasikan calon pengguna baru ke dalam segmen yang relevan segera setelah data mereka diterima. Ini akan mempercepat proses pengambilan keputusan dan personalisasi penawaran.
- **Integrasi ke Sistem Kredit dan Pemasaran:** Integrasikan model segmentasi ini secara langsung ke dalam sistem pengambilan keputusan di bagian kredit (untuk penilaian risiko dan penetapan limit) dan bagian pemasaran (untuk penargetan kampanye dan personalisasi komunikasi).

- **Monitoring dan Penyesuaian Berkelanjutan:** Lakukan monitoring berkala terhadap kinerja setiap segmen dan efektivitas strategi yang diterapkan. Data baru harus dianalisis secara periodik untuk menyesuaikan model *clustering* dan strategi bisnis agar tetap relevan dengan perubahan pasar dan perilaku pelanggan.

  Dengan mengadopsi strategi berbasis segmentasi yang terpersonalisasi ini, perusahaan kartu kredit tidak hanya dapat meningkatkan **efektivitas pemasaran** dan **tingkat konversi**, tetapi juga secara signifikan **menurunkan risiko kredit** melalui pengelolaan yang lebih cermat, serta **meningkatkan loyalitas dan kepuasan pelanggan** dalam jangka panjang. Pendekatan ini akan mendorong pertumbuhan bisnis yang lebih berkelanjutan dan menguntungkan.