
Customer Segmentation* berdasarkan Usia, Jumlah Kredit dan Lama Kredit Nasabah di Bank XYZ menggunakan Model *K-Means Clustering

Moch Rizky Wijaya¹, Gigih Satriyo Wibowo²

^{1,2}Jurusan Fisika, Universitas Brawijaya, Jalan Veteran, Ketawanggede, Lowokwaru, Malang, East Java 65145

Correspondence: Moch Rizky Wijaya (moch.wijayarizky@gmail.com)

Received: 01 08 21 – Revised: 04 08 21 - Accepted: 23 08 21 - Published: 26 10 21

Abstrak. Pelanggan internet banking tumbuh sangat cepat. Segmentasi nasabah dapat diterapkan berdasarkan data internet banking. *Clustering* adalah teknik *data mining* tanpa pengawasan yang dapat digunakan untuk segmentasi pelanggan. Penelitian ini membangun model clustering pada data profil nasabah bank berdasarkan data kredit sehingga didapatkan segmentasi nasabah yang nantinya digunakan sebagai landasan keputusan untuk melakukan strategi pemasaran. Metode clustering menggunakan metode K-Means dengan validasi *cluster* menggunakan metode *Silhouette coefficient*. Berdasarkan *Silhouette coefficient* didapatkan nilai terbaik untuk 3 *cluster* yaitu *cluster* 0, 1, dan 2. Hasil *cluster* dengan k-means terbagi menjadi 3 *cluster* yaitu *cluster* 0 – rata-rata jumlah kredit lebih rendah, durasi pendek, dan pelanggan usia tua, *cluster* 1 – rata-rata jumlah kredit tinggi, durasi panjang, dan pelanggan paruh baya atau usai pertengahan, dan *cluster* 2 - rata-rata jumlah kredit lebih rendah, durasi pendek, dan pelanggan usia muda. Hasil segmentasi tersebut dapat dijadikan acuan untuk melakukan strategi pemasaran kedepan.

Kata kunci: *Customer Segmentation, k-means clustering, data mining, bank*

Citation Format: Wijaya, M.R, Wibowo, G.S.. (2021). Customer Segmentation berdasarkan Usia, Jumlah Kredit dan Lama Kredit Nasabah di Bank XYZ menggunakan Model K-Means Clustering. *Prosiding Seminar Nasional Abdimas Ma Chung (SENAM)*, 2021, 101-116.

PENDAHULUAN

Seiring dengan perkembangan teknologi informasi, beberapa strategi bisnis termasuk bidang perbankan mulai melakukan *digitalisasi* bisnis dengan membuat layanan *electronic dealings* salah satunya adalah *internet banking*. *Internet banking* memudahkan nasabah bank dalam melakukan beberapa transaksi maupun pembelian beberapa produk, termasuk kredit. Tentunya setiap data transaksi yang dilakukan nasabah bank dapat dimanfaatkan untuk melakukan segmentasi pasar, mengingat jumlah nasabah yang menggunakan *internet banking* terus mengalami kenaikan setiap tahunnya (Bank Indonesia, 2003).

Segmentasi pasar merupakan proses aktual untuk mengidentifikasi atau membagi basis pelanggan yang luas menjadi sub-kelompok pelanggan berdasarkan variabel dari data pelanggan (Fyall et al., 2019). Segmentasi nasabah dapat diterapkan berdasarkan data nasabah dari *internet banking*. *Clustering* merupakan teknik *unsupervised data mining* yang dapat digunakan untuk melakukan segmentasi nasabah bank berdasarkan data demografi, jumlah kredit, dan lama kredit, sehingga menghasilkan segmentasi pasar nasabah yang sesuai target pemasaran (Wahyuni et al., 2016)

Tujuan dari data mining di sektor perbankan adalah untuk menganalisis retensi pelanggan, melakukan pencegahan penipuan dan deteksi dini penipuan (Chitra & Subashini, 2013). Data mining juga digunakan untuk segmentasi pelanggan yang melakukan kredit (Dhuhita, 2015). Segmentasi pelanggan akan menghasilkan segmen pelanggan yang masing-masing segmen memiliki karakteristik yang sama. Setiap cluster berisi titik-titik data dengan kemiripan yang tinggi tetapi berbeda secara signifikan dengan titik-titik data cluster lainnya (Sembiring Brahmana et al., 2020)

Penelitian serupa sebelumnya telah dilakukan oleh (Sembiring Brahmana et al., 2020) dengan melakukan proses segmentasi nasabah bank menggunakan k-means, k-medoids, dan DBSCAN. Hasil penelitian menunjukkan bahwa k-means memiliki tingkat validitas yang paling baik dibandingkan medoids dan DBSCAN, dimana yield Davies-Bouldin Index sebesar 0,33 dan yield silhouette index sebesar 0,9126. Jumlah cluster terbaik yang dihasilkan menggunakan Davies Bouldin Index dan Silhouette Index adalah 2 cluster, dimana masing-masing algoritma menyediakan kelas pelanggan *dormant* dan *golden*. Merujuk dari penelitian tersebut, pada penelitian ini berfokus menggunakan k-means untuk melakukan segmentasi data kredit nasabah bank dengan penentuan jumlah cluster terbaik menggunakan metode *silhouette*.

METODE PELAKSANAAN

Data set Penelitian

Pada penelitian ini menggunakan *data set* yaitu *german credit data* yang diakses dari UCI Center for Machine Learning and Intelligent Systems (<https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>). Berikut merupakan tampilan *data set* pada *output* perintah program python seperti yang ditunjukkan pada gambar 1.

	Unnamed: 0	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose
0	0	67	male	2	own	NaN	little	1169	6	radio/TV
1	1	22	female	2	own	little	moderate	5951	48	radio/TV
2	2	49	male	1	own	little	NaN	2096	12	education
3	3	45	male	2	free	little	little	7882	42	furniture/equipment
4	4	53	male	2	free	little	little	4870	24	car

Gambar 1. Data set Kredit Nasabah Bank Jerman

Sumber data : (<https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>)

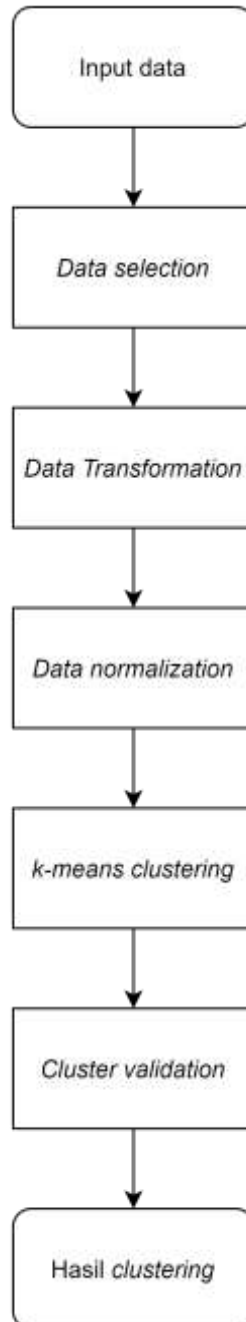
Ada 10 atribut yang terdapat pada data tersebut, berikut merupakan penjelasan setiap atribut data seperti yang ditunjukkan pada gambar 1.

1. *Age* merupakan data usia nasabah dengan format numerik
2. *Sex* merupakan data jenis kelamin nasabah dengan format *string*
3. *Job* merupakan data pekerjaan nasabah dengan format numerik dengan keterangan (0 : tidak memiliki keahlian dan bukan penduduk tetap, 1 : tidak memiliki keahlian dan merupakan penduduk, 2 : memiliki keahlian, dan 3 : sangat terampil)
4. *Housing* merupakan data kepemilikan rumah nasabah yaitu milik sendiri, sewa, atau gratis data ini merupakan data bertipe *string*
5. *Saving accounts* merupakan data rekening tabungan nasabah yaitu sedikit, sedang, cukup kaya, atau kaya dengan format data *string*
6. *Checking account* merupakan data rekening giro nasabah dengan format numerik dalam *Deustch Mark*
7. *Credit amount* merupakan data jumlah kredit nasabah dengan format tipe data numerik
8. *Duration* merupakan jangka waktu kredit nasabah dengan format tipe data numerik

9. *Purpose* merupakan tujuan pengajuan kredit dengan tipe format data *string* yaitu mobil, perabotan, radio/TV, perbaikan rumah, pendidikan, bisnis, dan liburan)

Atribut data yang dilakukan pengolahan segmentasi dengan k-means adalah *Age*, *Credit amount*, dan *Duration* dengan 3 kluster segmentasi.

Diagram Alir Penelitian



Gambar 2. Diagram Alir Penelitian

Pertama tentunya dilakukan eksplorasi data dengan menggunakan statistik deskriptif. Data kemudian diseleksi berdasarkan atribut data, pada penelitian ini digunakan atribut *Age*, *Credit amount*, dan *Duration* dengan 3 kluster segmentasi. Tipe data yang belum sesuai formatnya dilakukan transformasi tipe data, misalnya untuk atribut data *age* yang awalnya *string* dirubah menjadi numerik. Selanjutnya data dinormalisasi agar data lebih mudah dipahami. Normalisasi adalah bagian dari transformasi data yang digunakan untuk mengubah data menjadi nilai yang mudah dipahami. Normalisasi digunakan untuk meningkatkan akurasi proses perhitungan numerik dengan skala data dalam rentang 0 sampai 1. Pada kasus ini yang dilakukan normalisasi adalah distribusi data atribut *Age*, *Credit amount*, dan *Duration* dengan menggunakan fungsi logaritma natural agar didapatkan data yang terdistribusi normal. Data berikutnya dilakukan pengelompokan k-means dan dilakukan validasi *cluster* dengan menggunakan metode *silhouette*. Nilai *silhouette* yang didapatkan berada pada rentang -1 sampai 1, smekain mendekati 1 menunjukkan bahwa data tersebut semakin tepat berada di *cluster* yang dimaksud. *Silhouette coefficient* negative menunjukkan bahwa data tersebut tidak tepat berada dalam *cluster* yang dimaksud.

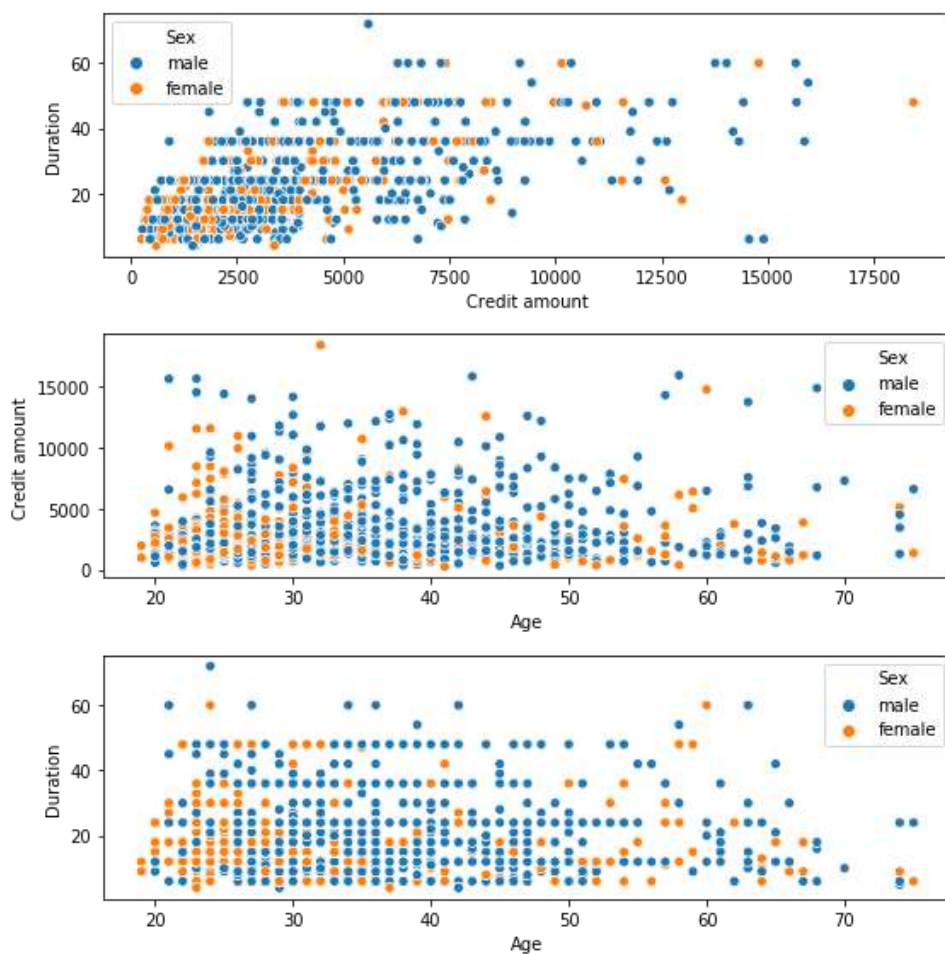
Pengelompokan K-Means

Dalam clustering K-Means, dataset, D, diberikan dari n titik data, dan k, jumlah cluster. Algoritma partisi mengatur titik-titik data ke dalam partisi (cluster), dimana $k < n$. Langkah-langkah dari K-means clustering adalah :

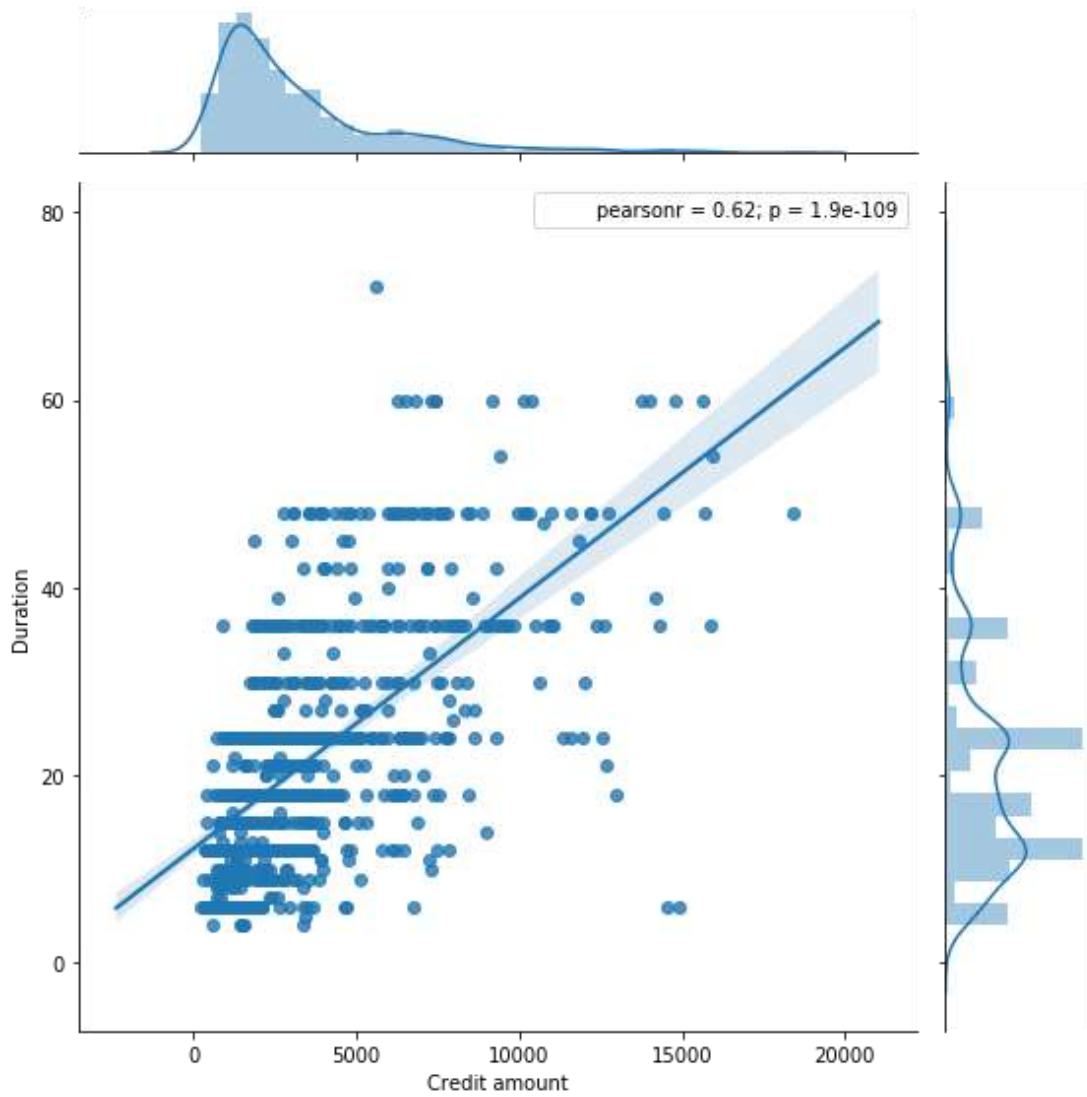
1. Tentukan jumlah cluster (k).
2. Pilih centroid awal.
3. Petakan setiap titik data ke dalam cluster terdekat (paling mirip dengan centroid).
4. Update nilai rata-rata (centroid) dari setiap cluster.
5. Ulangi langkah 3-4 sampai semua centroid tidak berubah.

HASIL DAN PEMBAHASAN

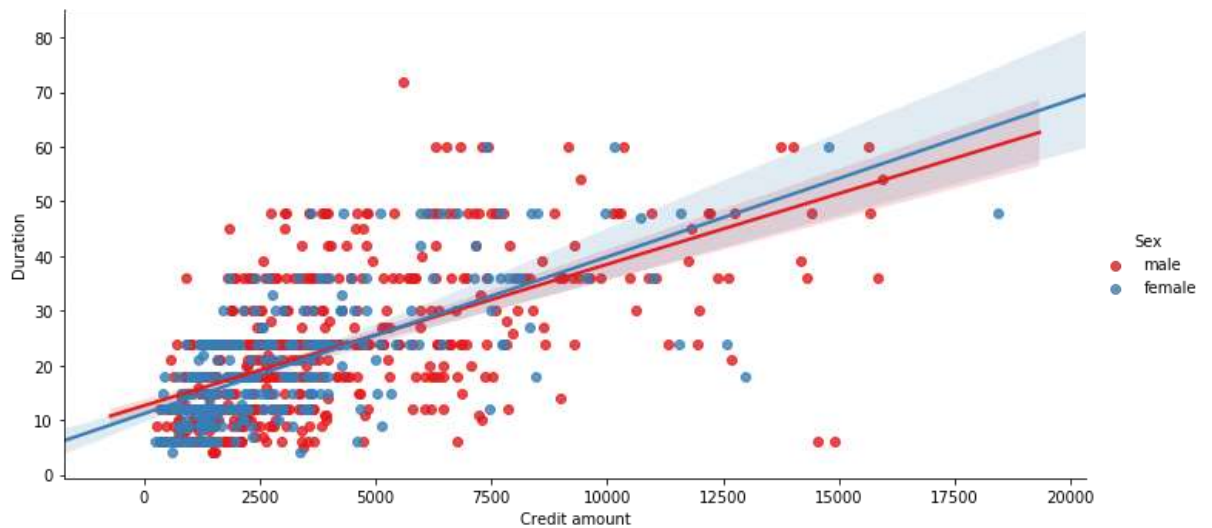
Hal pertama yang perlu dilakukan dalam analisis adalah melihat statistik deskriptif dan sebaran data nasabah bank yaitu berupa durasi transaksi, jumlah kredit, umur, dan jenis kelamin seperti yang ditunjukkan pada gambar 3. Pada gambar 3, menunjukkan bahwa nasabah perempuan cenderung lebih muda dari laki-laki, namun plot menunjukkan bahwa tidak ada perbedaan yang jelas antara laki-laki dan perempuan dalam hal jumlah dan durasi kredit. Dari pemeriksaan secara visual, kemungkinan ada korelasi positif antara durasi dan jumlah kredit. Sehingga dilakukan korelasi pearson antara atribut durasi kredit dan jumlah kredit seperti yang ditunjukkan gambar 4. Hasil korelasi seperti gambar 4, menunjukkan bahwa terjadi korelasi linier yang positif didukung dengan nilai Pearson sebesar 0,62 dan nilai p yang sangat kecil. Hal ini karena nasabah cenderung mengambil kredit yang lebih besar untuk waktu yang lebih lama.



Gambar 3. Scatter Plot Durasi transaksi, jumlah kredit, umur, dan Jenis kelamin dari Nasabah XYZ

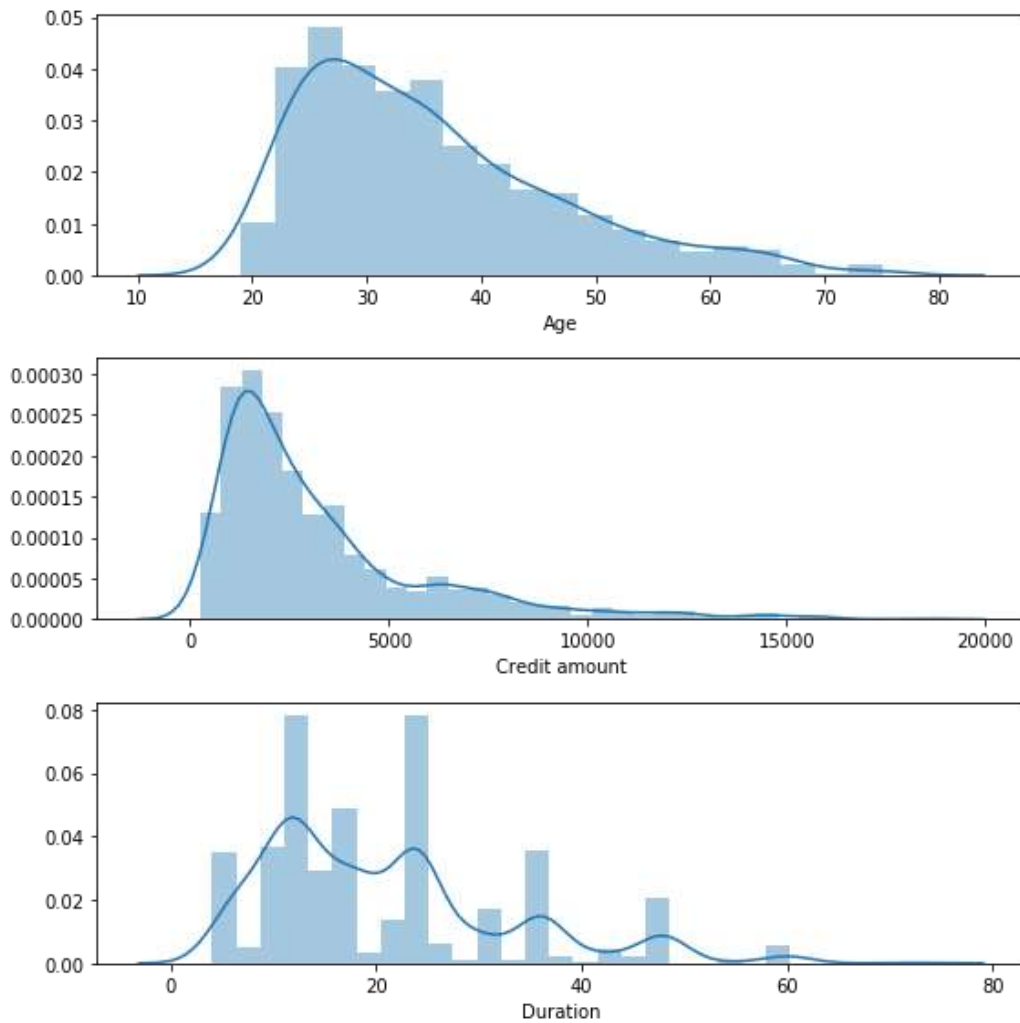


Gambar 4. Korelasi pearson antara jumlah kredit dan durasi kredit



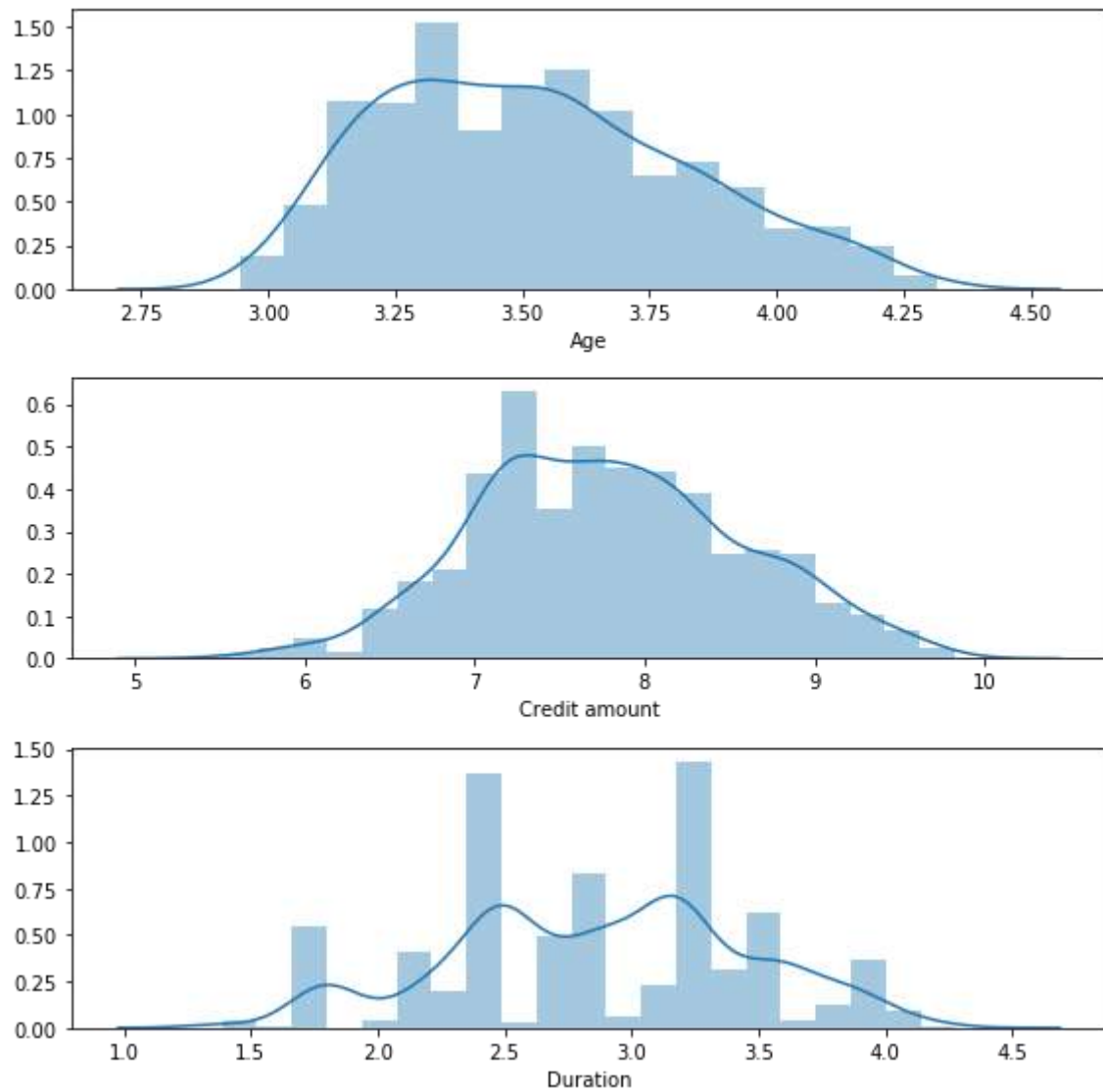
Gambar 5. Plot Durasi Kredit dan Jumlah Kredit berdasarkan Jenis Kelamin Nasabah Bank

Seperti yang ditunjukkan pada gambar 5, bahwa tidak ada perbedaan yang signifikan antara nasabah perempuan dan laki-laki dalam hubungan durasi kredit dan jumlah kredit. Sehingga selanjutnya dapat dilakukan *k-means clustering* berdasarkan atribut *Age*, *Credit amount*, dan *Duration* tanpa mempertimbangkan jenis kelamin/gender nasabah bank. Sebelum melakukan *clustering* terlebih dahulu data dilihat distribusi datanya, apakah data terdistribusi normal atau tidak seperti yang ditunjukkan pada gambar 6.



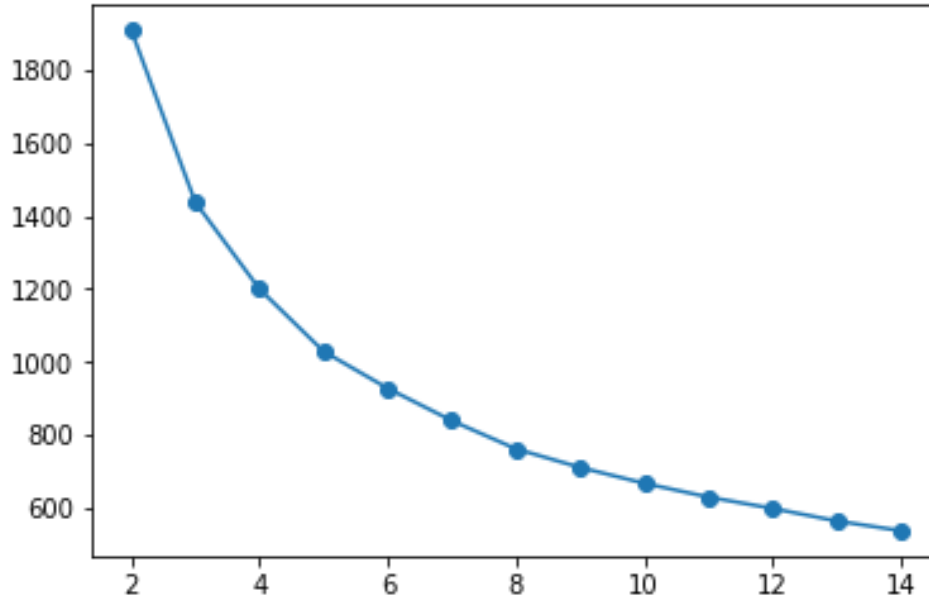
Gambar 6. Distribusi Data Nasabah Bank XYZ

Namun distribusi data menunjukkan bahwa distribusinya *right-skewed*. Sehingga untuk mendapatkan hasil yang lebih baik, harus menghilangkan skewness dengan transformasi logaritmik menggunakan fungsi “`np.log()`” pada python yaitu fungsi logaritma natural sehingga mendapatkan hasil distribusi data yang lebih terdistribusi normal seperti yang ditunjukkan pada gambar 7.

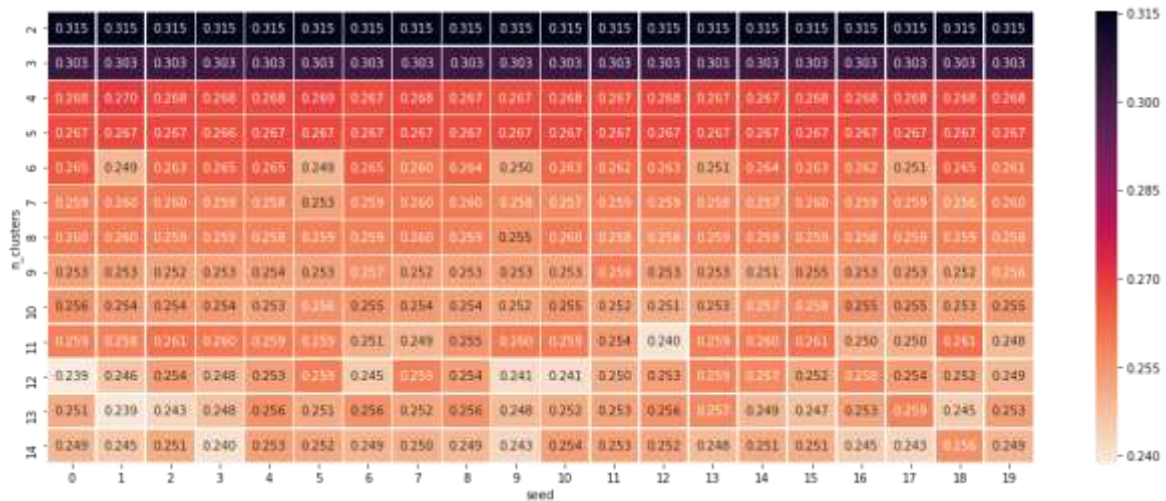


Gambar 7. Distribusi Data setelah dilakukan Transformasi Logaritma dengan Fungsi Logaritma Natural

Plot k-means pada gambar 8. menunjukkan bahwa inersia berkurang dengan bertambahnya jumlah cluster. Metode ini memungkinkan penilaian pemisahan kluster dan pemasangan setiap pengamatan di klasternya sendiri. Skor tertinggi menunjukkan hasil yang lebih baik.

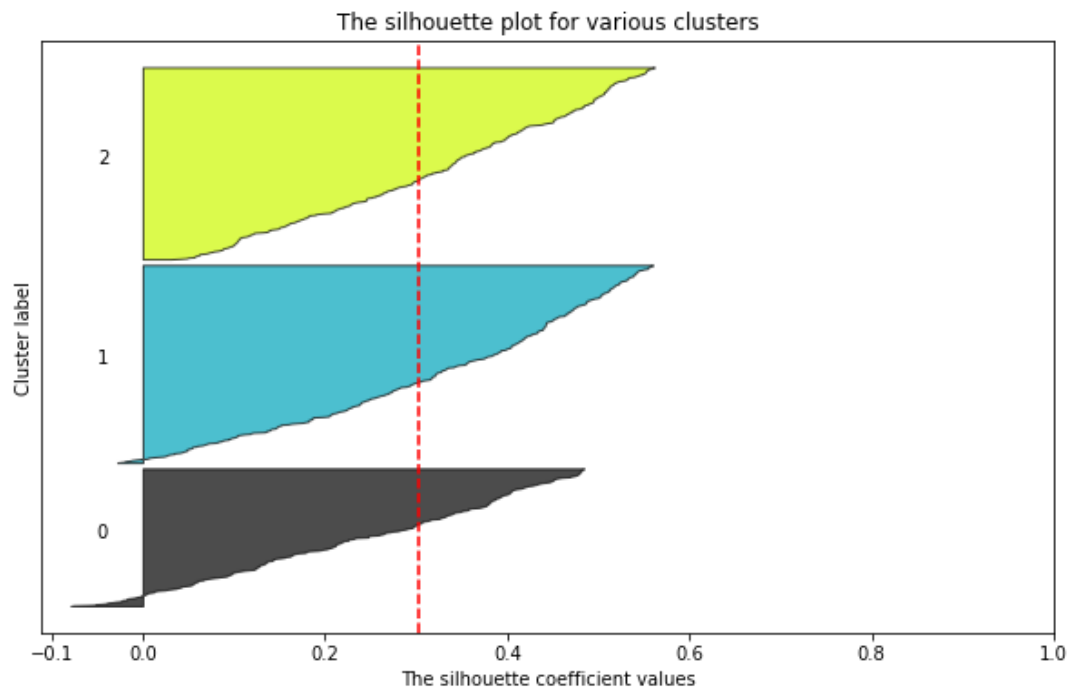


Gambar 8. Hasil K-Means Clustering



Gambar 9. Skor Silhouette coefficient untuk berbagai kombinasi keadaan acak dan jumlah cluster

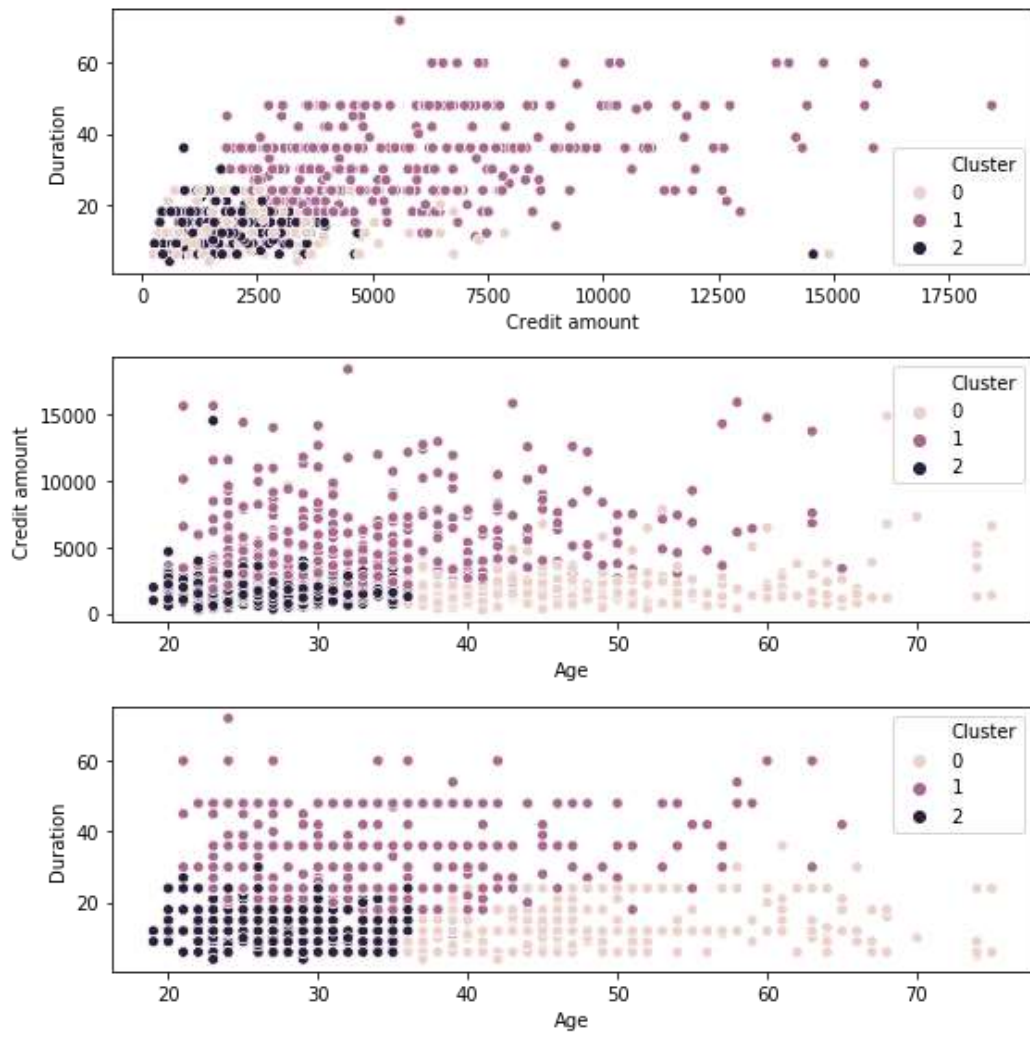
Heatmap seperti gambar 9 menunjukkan skor Silhouette coefficient untuk berbagai kombinasi keadaan acak dan jumlah cluster. Skor tertinggi untuk cluster 2 dan 3 dan relatif tidak sensitif terhadap seed. Sehingga berdasarkan Silhouette coefficient dipilih 3 cluster (cluster 0, 1, dan 2) untuk mendapatkan hasil cluster yang lebih baik. Terkait rata-rata nilai Silhouette coefficient untuk 3 cluster dapat dilihat pada gambar 10.



Gambar 10. Hasil Perhitungan Rata-Rata Silhouette untuk Variasi Cluster

Berdasarkan hasil plot k-means clustering seperti pada gambar 11 dan hasil rata-rata nilai setiap *cluster* seperti pada gambar 12 menunjukkan bahwa:

1. Cluster 0 – rata-rata jumlah kredit lebih rendah, durasi pendek, pelanggan usia paling tua
2. Cluster 1 – rata-rata jumlah kredit tinggi, durasi panjang, pelanggan usia pertengahan
3. Cluster 2 - rata-rata jumlah kredit lebih rendah, durasi pendek, pelanggan usia muda



Gambar 11. Hasil Plot Clustering dengan K Means

Hasil ketiga *cluster* seperti pada gambar 12, menunjukkan bahwa *Cluster 0* – rata-rata jumlah kredit lebih rendah, durasi pendek, dan pelanggan usia paling tua, hal ini karena pada usia 45 tahun ke atas nasabah bank lebih melakukan pinjaman yang sedikit mengingat kebutuhannya tidak terlalu banyak, karena diusia tersebut kemungkinan anaknya sudah menyelesaikan pendidikan dan beban orang tua berkurang. Sedangkan pada *Cluster 1* – rata-rata jumlah kredit tinggi, durasi panjang, dan pelanggan usia pertengahan dimana pada usia 30 – 40 tahun, nasabah mulai mempunyai anak dan merencanakan untuk mempunyai rumah sendiri dengan pasangannya sehingga kemungkinan memutuskan untuk melakukan pinjaman dengan jumlah yang tinggi dengan durasi pengembalian yang panjang. Disisi lain, *Cluster 2* - rata-rata jumlah kredit lebih rendah, durasi pendek, pelanggan usia muda, dimana pada usia tersebut sekitar 20 – 28 tahun, nasabah bank masih belum memutuskan menikah dan masih belum banyak tanggungan sehingga lebih memutuskan untuk melakukan kredit yang kecil dengan jatuh tempo pengembalian yang singkat.

Hasil *clustering* ini dapat dijadikan acuan untuk melakukan startegi pemasaran yang sesuai dengan target pasar yang dituju, tetutama untuk memasarkan produk kredit bank ke nasabah.

	Age	Credit amount	Duration
Cluster			
0	48.6	1970.5	13.9
1	34.0	5665.4	32.2
2	27.7	1737.5	14.3

Gambar 12. Hasil Rata-Rata Nilai untuk Setiap Cluster

KESIMPULAN

Penelitian ini membangun model clustering pada data profil nasabah berdasarkan penggunaan data kredit untuk segmentasi nasabah di bank XYZ menggunakan metode K-Means. Hasil *clustering* yang dilakukan dengan menggunakan metode k-means menunjukkan sebagai berikut:

1. Cluster 0 – rata-rata jumlah kredit lebih rendah, durasi pendek, dan pelanggan usia paling tua
2. Cluster 1 – rata-rata jumlah kredit tinggi, durasi panjang, pelanggan usia pertengahan

3. Cluster 2 - rata-rata jumlah kredit lebih rendah, durasi pendek, pelanggan usia muda

Saran untuk penelitian kedepan: (1) *clustering* diterapkan dengan menggabungkan data transaksi Internet Banking, informasi sosio-demografis dan data kepemilikan produk, dan (2) membuat aturan perilaku nasabah menggunakan metode klasifikasi..

UCAPAN TERIMA KASIH

Ucapan terima kasih ditujukan kepada semua pihak yang terlibat dalam pembuatan naskah. Ucapan terima kasih dapat pula ditujukan kepada penyedia dana atau sponsor.

DAFTAR PUSTAKA

- Bank Indonesia, T. P. P. P. (2003). Internet Banking Di Indonesia. *Buletin Ekonomi Moneter Dan Perbankan*, 5(1), 37–64. <https://doi.org/10.21098/bemp.v5i1.304>
- Chitra, K., & Subashini, B. (2013). Data Mining Techniques and its Applications in Banking Sector. *International Journal of Emerging Technology and ...*, 3(8), 219–226. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Data+Mining+Techniques+and+its+Applications+in+Banking+Sector#0>
- Duhita, W. (2015). Clustering Menggunakan Metode K-Mean Untuk Menentukan Status Gizi Balita. *Jurnal Informatika Darmajaya*, 15(2), 160–174.
- Fyall, A., Legoh erel, P., Frochot, I., & Wang, Y. (2019). Market segmentation, targeting, and positioning. *Marketing for Tourism and Hospitality*, 321–342. <https://doi.org/10.4324/9781315651026-25>
- Sembiring Brahmana, R. W., Mohammed, F. A., & Chairuang, K. (2020). Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, 11(1), 32. <https://doi.org/10.24843/lkjiti.2020.v11.i01.p04>
- Wahyuni, I., Auliya, Y. A., Rahmi, A., & Mahmudy, W. F. (2016). Clustering Nasabah Bank Berdasarkan Tingkat Likuiditas Menggunakan Hybrid PSO K-Means. *Jurnal Ilmiah Teknologi Dan Informasi ASIA (JITIKA)*, 10(1), 24–33.

