

Penerapan *Clustering* Pada Tipe Data Campuran Menggunakan *K-Prototype* Pada Perusahaan *Multifinance*

1st Oktavianus Jeffry Pradhana
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia

oktavianusjeffry@students.telkomuni-
versity.ac.id

2nd Deni Saepudin
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia

denisaepudin@telkomuniversity.ac.id

3rd Indwiarti
Fakultas Informatika
Universitas Telkom
Bandung, Indonesia

indwiarti@telkomuniversity.ac.id

Abstrak

Multifinance merupakan perusahaan pembiayaan keuangan yang berperan dalam penyediaan dana nasabah untuk pengadaan barang berdasarkan kebutuhan produktif maupun konsumtif masyarakat. Dalam menjalankan operasionalnya perusahaan melakukan observasi data nasabah berdasarkan *profiling* resiko nasabah untuk menentukan seberapa banyak pinjaman yang dapat diberikan oleh perusahaan. Namun data nasabah yang terdiri dari banyak numerik dan kategori serta memiliki bentuk yang tidak terstruktur membuat lembaga pembiayaan kesulitan dalam memberikan keputusan pada proses pembiayaan. Maka diperlukan alat yang dapat melakukan pengelompokan data nasabah berdasarkan tingkat kesamaan atau kedekatan antar data sehingga mempermudah perusahaan dalam melakukan segmentasi profil nasabah berdasarkan tingkat resiko. Untuk mengatasi masalah tersebut maka pada penelitian ini menggunakan metode *Clustering* untuk mengelompokan data tipe campuran pada data nasabah perusahaan *multifinance* PT Bima *Multifinance* Cabang Sragen menggunakan algoritma *k-prototype*. Prosedur kerja pada algoritma *k-prototype* dengan mengintegrasikan algoritma *k-means* dan *k-modes* yang digunakan untuk menangani tipe data campuran. Selain itu algoritma *k-prototype* memiliki keunggulan seperti mudah untuk diterapkan dan serta mampu untuk menangani kumpulan data besar lebih baik dibandingkan algoritma yang berbasis hirarki lain. Pengujian pada penelitian ini menggunakan perhitungan elbow untuk mencari *cluster* optimal dan melakukan analisis lanjutan dengan melakukan pengujian ulang pada tahap inisiasi jumlah *cluster*. Hasil *clustering* pada penelitian ini menunjukkan salah satu jenis *cluster* dengan tingkat kelancaran 100% dan jumlah rasio gaji terhadap angsuran nasabah dibawah rata rata yaitu 1.3% yang membuat *cluster* ini menjadi *cluster* dengan tingkat kelancaran tertinggi dibanding *cluster* lain.

Kata kunci : *Multifinance*, Data Tipe Campuran, *Clustering*, *K-Prototype*, *K-Means*, *K-Modes*.

Abstract

Multifinance is a financial finance company that plays a role in providing customer funds for procurement of goods based on the productive and consumptive needs of the community. In carrying out its operations, the Company observes customer data based on customer risk profiling to determine how much loan the company can provide. However, customer data which consists of many numbers and categories and has an unstructured form makes it difficult for financial institutions to make decisions on the financing process. then we need a tool that can group customer data based on the level of similarity or proximity between the data so as to make it easier for companies to segment customer profiles based on the level of risk. To overcome this problem, this study uses the *Clustering* method to group mixed-type data on customer data of a multi-finance company PT Bima *Multifinance* Branch Sragen using the *k-prototype* algorithm. The working procedure of the *k-prototype* algorithm is by integrating the *k-means* and *k-modes* algorithms that are used to handle mixed data types. In addition, the *k-prototype* algorithm has advantages such as being easy to implement and being able to handle large data sets better than other hierarchical-based algorithms. The test in this study uses elbow calculations to find the optimal cluster and performs further analysis by re-testing at the initiation stage of the number of clusters. The results of clustering in this study show one type of cluster with a 100% smoothness rate and the total salary to customer installment ratio is below the average, namely 1.3% which makes this cluster a cluster with the highest smoothness rate compared to other clusters.

Keywords: *Multifinance*, Mixed Type Data, *Clustering*, *K-Prototype*, *K-Means*, *K-Modes*.

I. PENDAHULUAN

Perusahaan pembiayaan atau *Multifinance* merupakan salah satu Lembaga keuangan bukan Bank di Indonesia, yang berperan dalam penyediaan dana nasabah untuk pengadaan barang berdasarkan kebutuhan produktif maupun konsumtif masyarakat[1]. Salah satu perusahaan pembiayaan yang bergerak dalam bidang ini

adalah PT Bima *Multifinance* Cabang Sragen. Dalam menjalankan operasionalnya perusahaan dituntut harus mampu memenuhi target keuntungan dengan menghindari resiko nasabah yang gagal bayar serta memastikan semua proses pembiayaan dilakukan sesuai pedoman bisnis yang sehat, mengingat setiap pegawai dibebani target yang besar bisa saja beberapa pegawai tidak peduli dengan prosedur ini dan memberikan pembiayaan yang salah kepada nasabah[2]. Untuk itu perusahaan perlu mengumpulkan data nasabah terlebih dahulu seperti fotocopy KTP, BPKB, STNK dan beberapa sejenisnya. Perusahaan lalu mengolah data nasabah berdasarkan beragam kriteria seperti informasi usia, jenis kelamin, pekerjaan, gaji, jaminan dan sebagainya. Kemudian dilakukan proses segmentasi *profiling* nasabah dalam menentukan seberapa banyak pinjaman yang dapat diberikan oleh perusahaan. Namun data nasabah yang terdiri dari banyak numerik dan kategori serta memiliki bentuk yang tidak terstruktur membuat lembaga pembiayaan kesulitan untuk menganalisa dan memberikan keputusan pada proses pembiayaan. maka diperlukan alat yang dapat melakukan pengelompokan data nasabah berdasarkan tingkat kesamaan atau kedekatan antar data sehingga mempermudah perusahaan dalam melakukan segmentasi profil nasabah berdasarkan tingkat resiko.

Berdasarkan penelitian berjudul "*Integration Clustering with K-Means algorithm and Elbow Method For Identification of The Best Customer Profile*" pada tahun 2018 metode *clustering* mampu digunakan untuk melakukan pengelompokan pada data nasabah berdasarkan kriteria secara optimal(Syakur, M. A., 2018). Penerapan *clustering* pada penelitian tersebut menggunakan algoritma *k-means* dengan menginisiasi *centroid* secara acak lalu menghitung jarak kemiripan antar data pada atribut yang bersifat numerik[8]. Namun menurut study penelitian terkait *clustering* pada tahun 2007, algoritma *k-means* memiliki kekurangan dalam menghadirkan kumpulan data pada penetapan *cluster* awal yang dipilih secara acak terutama untuk menangani atribut kategori pada perusahaan kredit. Untuk mengatasi masalah pada data tipe campuran diperlukan

algoritma lain yaitu *k-prototype*, dimana prosedur kerja pada algoritma *k-prototype* mengintegrasikan algoritma *k-modes* yang digunakan untuk menangani atribut kategori dan algoritma *k-means* untuk menangani atribut numerik, lalu mengelompokkan kumpulan data tersebut dengan menggabungkan hasil perhitungan. Selain metode pengelompokan *k-prototype* yang termasuk baru, algoritma *k-prototype* memiliki keunggulan seperti mudah untuk diimplementasikan dan algoritma ini mampu untuk menangani kumpulan data besar lebih baik dibandingkan algoritma yang berbasis hirarki lain. [4][5]

Berdasarkan penjabaran terkait masalah di atas, maka dalam penelitian memutuskan menggunakan metode *clustering* pada data tipe campuran dengan algoritma *k-prototype* pada data perusahaan Bima *Multifinance* cabang Sragen. Pemilihan metode ini bertujuan sebagai bahan pertimbangan perusahaan dalam mempermudah perusahaan melakukan segmentasi profil nasabah berdasarkan tingkat resiko.

II. KAJIAN TEORI

A. Penelitian terkait

Dalam penelitian ini penulis mengambil referensi dari penelitian terkait yang menggunakan metode *clustering* pada data tipe campuran contohnya pada penelitian yang berjudul "*Integration Clustering with K-Means algorithm and Elbow Method For Identification of Customer Profile*" yang dilakukan oleh M syakur pada tahun 2018. Dalam penelitian menerapkan kombinasi metode *K-Means* dengan *elbow* untuk meningkatkan kinerja *clustering k-means* untuk melakukan segmentasi pelanggan dalam jumlah besar. Menurut hasil penelitian didapatkan penentuan jumlah *cluster* terbaik menggunakan gabungan metode *elbow* pada jumlah data yang berbeda dapat menghasilkan jumlah *cluster k* yang sama. Lalu hasil dari penentuan jumlah *cluster* terbaik dengan metode *elbow* menjadi parameter dalam proses penentuan banyak *cluster profiling* pelanggan berdasarkan studi kasus[8]. Selanjutnya pada penelitian yang berjudul "*Applications of Clustering with Mixed Type Data in Life Insurance*" yang dilakukan oleh Shuang yin pada tahun 2021. Penelitian ini meneliti bagaimana klaim kematian pada data nasabah laporan keuangan daerah di amerika pada perusahaan asuransi jiwa. Metode yang digunakan pada penelitian ini yaitu *clustering* menggunakan *k-prototype*. Metode ini mampu untuk menangani tipe kategorikal dan numerik. lalu dari hasil pengelompokan diperoleh *cluster* optimal yang kemudian digunakan untuk

membandingkan pengalaman klaim kematian secara aktual dengan yang diharapkan dari portfolio perusahaan asuransi jiwa. Hasil penelitian menyimpulkan bahwa proses manajemen dalam atribut pemegang polis mendominasi secara signifikan pada penyimpangan kematian, dan dengan demikian meningkatkan efisiensi pengambilan keputusan dalam mengambil tindakan yang diperlukan pada klaim kematian[3]. Kemudian penelitian terakhir yang berjudul “Clustering Pada Data Campuran K-Prototype Untuk Menangkap Karakteristik Pasien Gagal Jantung Berdasarkan Variabel karakteristik” yang dilakukan oleh Raditya novidianto dan Kartika fithriasari pada tahun 2021. Pada penelitian ini meneliti tentang pendeteksian faktor mortalitas pada pasien gagal jantung untuk memperkecil peluang terjadinya kematian akibat gagal jantung dengan metode *clustering* menggunakan algoritma *k-prototype*. Penelitian ini mengelompokkan 2 *cluster* yang dianggap optimal berdasarkan nilai koefisien *silhouette* tertinggi 0,5777. Lalu hasil dari penelitian menunjukkan bahwa *cluster* 1 menunjukkan *profil cluster* yang memiliki resiko rendah terhadap gagal jantung dan sedangkan pada *cluster* 2 merupakan *profil*

cluster dengan resiko yang tinggi terhadap peluang gagal jantung. Data tersebut diambil dari nilai rata-rata setiap variabel penciri dan karakteristik mortalitas pasien gagal jantung [7].

B. Multifinance

Perusahaan pembiayaan atau *Multifinance* merupakan salah satu Lembaga keuangan bukan Bank di Indonesia, yang berperan dalam penyediaan dana nasabah untuk pengadaan barang berdasarkan kebutuhan produktif maupun konsumtif masyarakat. Salah satu perusahaan pembiayaan yang bergerak dalam bidang ini adalah PT Bima *Multifinance* Cabang Sragen. Dalam menjalankan operasional perusahaannya manajemen perusahaan dituntut harus mampu memenuhi target keuntungan perusahaan, dalam artian lain keuntungan maksimal dengan menghindari resiko nasabah yang gagal bayar serta memastikan semua proses pembiayaan dilakukan sesuai pedoman bisnis yang sehat, mengingat setiap pegawai dibebani target yang besar bisa saja beberapa pegawai tidak peduli dengan prosedur ini dan memberikan pembiayaan yang salah kepada nasabah[1].

C. Clustering

Clustering merupakan metode *unsupervised* yang digunakan untuk mengelompokkan data yang memiliki karakteristik kemiripan dengan satu data dengan data yang lain. Metode

Penerapan pada metode ini tanpa adanya latihan dan tanpa adanya pengajar, serta tidak memerlukan target *output*. *Clustering* memiliki dua jenis pengelompokan data, yaitu *hierarchical* dan *non-hierarchical clustering*. Pada pengelompokan *hierarchical* dimulai dengan mengelompokkan dua atau lebih objek berdasarkan *similarity measure*. Kemudian proses akan diteruskan ke objek lain yang memiliki kedekatan kedua hingga *cluster* akan membentuk semacam tingkatan layaknya sebuah pohon, dari tingkat yang paling mirip hingga tidak mirip. Sedangkan pada *non-hierarchical clustering* dimulai dengan melakukan inisiasi *cluster* awal, proses selanjutnya dilakukan tanpa mengikuti proses hierarki[11].

D. Algoritma K-Prototype

Algoritma *k-prototype* termasuk dalam metode *nonhierarchical clustering* dan pertama kali dikemukakan oleh Huang (1998). Algoritma ini merupakan algoritma yang digunakan dalam *clustering* tipe data campuran (numerik dan kategori) dengan mengintegrasikan perhitungan jarak kemiripan pada algoritma *k-means* dan *k-modes*. Pada perhitungan jarak algoritma *k-prototype* menggunakan ukuran kesamaan antar objek dengan menggabungkan persamaan *euclidean distance* dengan *dissimilarity measure* yang ada dalam *k-modes*[9]. Untuk menggambarkan cara kerja algoritma *k-prototype*, misalkan $\{x_{ij}\}$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, d$ menunjukkan kumpulan data yang berisi n batas data pengamatan. Setiap observasi dideskripsikan oleh D variabel. Kemudian untuk *dissimilarity measure* antara dua titik \mathbf{x} dan \mathbf{y} yang digunakan pada algoritma *k-prototype* yang didefinisikan sebagai berikut:

$$D(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{d_1} (x_j - y_j)^2 + \lambda \sum_{j=d_1+1}^d \delta(x_j, y_j) \quad (1)$$

Dimana d_1 merupakan batas ukuran jarak untuk atribut bertipe numerik, sedangkan d merupakan batas ukuran jarak pada atribut kategorik. Sedangkan pada lambda λ merupakan parameter penimbang yang digunakan dalam menyeimbangkan perhitungan dua fungsi jarak kemiripan pada dua tipe data numerik dan kategorik. selanjutnya pada $\delta(x_j, y_j)$ adalah jarak pencocokan sederhana pada variabel kategorik yang didefinisikan pada persamaan (2)

$$\delta(x_j, y_j) = \begin{cases} 1, & \text{jika } x_j \neq y_j, \\ 0 & \text{jika } x_j = y_j, \end{cases} \quad (2)$$

Untuk inisiasi nilai lambda (λ) menggunakan

perhitungan nilai default yang sudah ditetapkan pada library dengan persamaan berikut.

$$\lambda = 0.5 \times std(X_{numerical}) \tag{3}$$

$std(X_{numerical})$ merupakan standar deviasi dari keragaman data berdasarkan atribut numerik yang terdapat dalam data yang digunakan pada pengukuran jarak kemiripan dua jenis data.

Algoritma k-prototype menggunakan perhitungan fungsi *cost*. Dimana *cost* didapatkan dari nilai minimal perhitungan jarak

$$\min(cost_{cluster\ 1}, cost_{cluster\ 2}, \dots, cost_{cluster\ n}) \tag{4}$$

kemiripan indeks *cluster* dengan pusat *cluster* didefinisikan pada persamaan (4).

Sedangkan untuk perhitungan total *cost* pada *clustering* dilakukan dengan persamaan berikut.

$$\sum_{j=1}^j \sum_{x=1}^x cost_{x,i} \tag{5}$$

Dimana :

- i* = indeks untuk identifikasi *clustering* anggota ke-*i*,
- j* = batas indeks anggota *cluster* untuk identifikasi nomor *cluster* terbesar.
- x* = indeks untuk mengidentifikasi baris anggota *cluster* dari tiap data,
- y* = baris terakhir pada indeks terakhir $cost_{x,i}$ = fungsi biaya pada baris ke-*x* dan *cluster* ke-*i*.

untuk proses pembaharuan *centroid* didapatkan dengan cara mencari modus dari setiap atribut katagorik dan mencari rata-rata (*means*) dari atribut numerik sehingga *centroid* pada tiap *cluster* akan berubah hingga kondisi konvergen sampai batas iterasi dilakukan[11]. Implementasi code untuk proses *generate cluster* pada tahap pembaruan *centroid* direpresentasikan sebagai berikut.

```
for cluster in range(1,n_cluster):
    kproto = KPrototypes(n_clusters=cluster,
        verbose=2,max_iter=10)
    kproto.fit_predict(mark_array,
        categorical=[0,2,7])
    cost.append(kproto.cost_)
    count_cluster.append(cluster)
```

dimana *KPrototypes* digunakan dalam inisiasi jumlah *cluster*, dan seberapa banyak iterasi dilakukan untuk mencapai titik *centroid*. Lalu pada *kproto.fit_predict* adalah proses melatih dan memprediksi titik *centroid* dengan mencari perhitungan *cost* minimal. Sedangkan pada *cost.append* memasukan nilai *cost* terbaik dari setiap iterasi untuk kemudian dilakukan

klasifikasi *cluster* sesuai titik *centroid* terdekat pada proses pemberian label *count_cluster.append*.

E. Metode Elbow

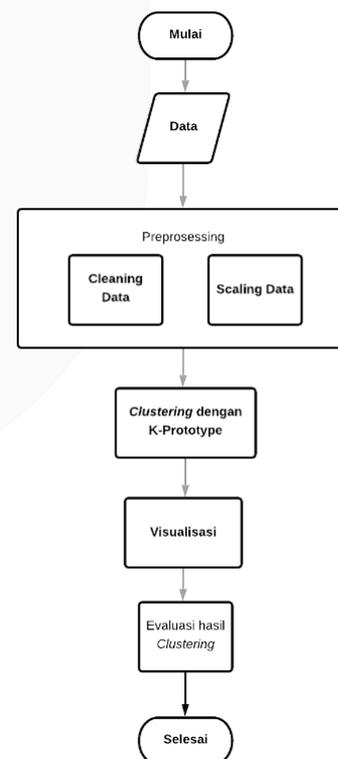
Metode Elbow merupakan metode yang digunakan dalam menentukan jumlah *cluster* optimal melalui perhitungan fungsi *cost*. Perhitungan fungsi *cost* pada metode elbow menggunakan persamaan berikut :

$$P(W, Q) = \sum_{i=1}^k \sum_{l=1}^n w_{i,l} \sum_{j=1}^p (x_{i,j} - q_{i,j}) + \gamma \sum_{j=1}^p w_{i,l} \sum_{j=p+1}^m \delta(x_{i,j}, q_{i,j})$$

Hasil dari fungsi *cost* pada metode elbow disebut inerti, dimana nilai inerti semakin rendah nilai inerti maka jarak tiap titik pusat *cluster* terhadap titik data semakin rendah. Jumlah *cluster* optimal didapatkan jika hasil penurunan inerti membentuk siku siku[8].

III. METODE

Pada penelitian ini, Peneliti akan membangun Perancangan Sistem untuk penerapan *clustering* pada tipe data campuran menggunakan *k-prototype* pada perusahaan *Multifinance* yang disajikan pada *flowchart* dibawah ini



GAMBAR 1. Flowchart Sistem

A. Pengumpulan Data

Data yang digunakan untuk penelitian ini merupakan data laporan nasabah tanggal 08 April 2021 sejumlah 1265 record yang dikumpulkan PT Bima *Multifinance* cabang Sragen. Dalam penelitian ini terdapat 9 variabel dari penciri faktor yang mempengaruhi tingkat resiko nasabah kredit berdasarkan hasil klaim perusahaan yang meliputi *Cust occu*, *Salary*, *BPKB Name*, *Harga Motor*, *Spouse Salary*, *Birth date*, *Angsuran* dan *Group overdue days* (Status kelancaran nasabah) . Rincian dataset dapat dilihat pada tabel 1.

TABEL 1. Deskripsi dataset

Nama Kolom	Tipe Data	Deskripsi
<i>Cust occu</i>	Numerik	Kolom berisi jenis pekerjaan nasabah
<i>Salary</i>	Numerik	Kolom berisi nominal gaji nasabah
<i>BPKB Name</i>	katagori kal	Kolom berisi status kepemilikan dari surat BPKB nasabah
Harga motor	Numerik	Kolom berisi harga objek yang menjadi jaminan nasabah
<i>Spouse salary</i>	Numerik	Kolom berisi nominal gaji penjamin nasabah
usia	Numerik	Kolom berisi tanggal lahir nasabah
Angsuran	Numerik	Kolom berisikan nominal angsuran kredit nasabah
<i>Group overdue days</i>	katagori kal	Kolom berisi detail dari keterlambatan kredit nasabah dengan beberapa jenis kategori

B. Preprocessing Data

Pada tahap preprocessing dilakukan eksplorasi data dengan beberapa tahapan yaitu *cleaning* data untuk menyiapkan data agar bebas dari *missing value* dan *scaling* data untuk mengurangi variasi pada data. Pada *cleaning* data dilakukan dengan menghilangkan data yang memiliki *missing value* atau mengganti data

yang kosong dengan nilai 0 agar sesuai dengan proses *clustering* tanpa menghilangkan nilai dari data tersebut. Sedangkan pada *scaling* data dilakukan untuk mengurangi katagori *unique* terlalu banyak pada data dengan mengelompokan jenis pekerjaan nasabah pada kolom *cust occu*.

TABEL 2. Scaling Data

Kategori	Pekerjaan
Lain-lain	Lain-Lain, Seniman, Produksi, Pendidikan Non Formal, Pengrajin Tangan, Ibu Rumah Tangga, Dokter/Bidan/Mantri, Pelajar/ Mahasiswa, Kesehatan, Nelayan
Pegawai swasta	Buruh/Prt, Karyawan Swasta, Sopir/Pengemudi, Transportasi & Komunikasi, Jasa
Wiraswasta	Pedagang, Sewa/Rent, Peternak, Petani, Wiraswasta
PNS	Pegawai Negeri Sipil, Guru/Pendidikan, Pegawai Negeri/ Bumh, Aparatur Pemerintah, Dosen, Pensiunan/ Purnawirawan Dan Penegak Hukum

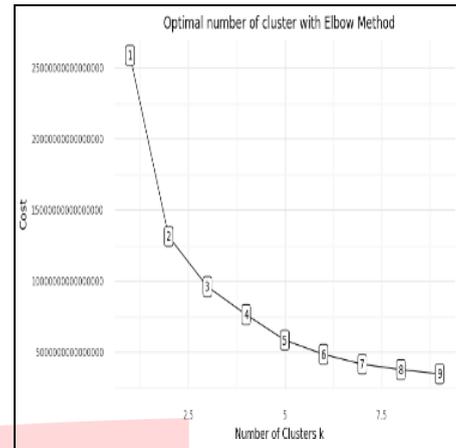
C. Clustering Algoritma K-Prototype

Pada tahap pengujian algoritma k-prototype penelitian ini menggunakan *library open source* yang tersedia pada google collab menggunakan bahasa python. Library ini dibangun dengan menerapkan penelitian Huang tentang algoritma *k-prototype*. Implementasi algoritma *k-prototype* dilakukan pada data sampel berjumlah 1260 record. Penentuan jumlah *cluster* yang optimal dilakukan menginisiasi nilai *k* menggunakan perhitungan *elbow* dengan memasukan nilai $k = 2$ hingga $k = 10$. Selanjutnya jika nilai optimal *cluster k* sudah ditentukan, maka proses *clustering* algoritma *k-prototype* dapat dilakukan dengan mengikuti tahapan sebagai berikut :

Langkah 1: Tentukan centroid *cluster* sebanyak *k cluster* titik awal $\{c_1, c_2, \dots, c_n\}$ pada setiap variabel $\{x_1, x_2, \dots, x_n\}$;

Langkah 2: Hitung jarak titik data pada kumpulan data terhadap *centroid cluster*, kemudian alokasikan titik data anggota *cluster* yang memiliki jarak

- terdekat dengan *centroid*;
- Langkah 3: Hitung *centroid* baru dari *cluster* setelah semua objek telah dialokasikan ke dalam *cluster*, dan kemudian realokasi semua objek pada prototipe baru;
- Langkah 4: Jika *centroid cluster* tidak berubah atau telah konvergen, maka algoritma akan berhenti. Namun, jika *centroid* masih berubah, proses harus kembali ke langkah 2 dan 3 sampai tidak ada perubahan pada objek.



GAMBAR 1 Perhitungan nilai k optimal elbow

Jika proses *clustering* menggunakan nilai optimal menunjukkan *cluster* yang homogen maka dilakukan pengujian lanjutan dengan melakukan pengujian ulang dari nilai $k = 2$. Nilai k akan bertambah seiring dengan percobaan yang dilakukan hingga batas pengujian nilai $k = 6$. Setelah itu melakukan perbandingan hasil dari percobaan tersebut untuk mencari hasil percobaan *clustering* yang menghasilkan *cluster* konvergen.

Pada Gambar 1 menunjukkan adanya perubahan nilai signifikan yang membentuk siku pada nilai $k = 1$ dengan $k = 2$ dibanding nilai k yang lain, sehingga nilai $k = 2$ ditetapkan sebagai jumlah *cluster* yang optimal. Hasil dari pengujian *cluster 2* dapat dilihat dibawah.

D. Visualisasi

Proses visualisasi merupakan tahapan setelah hasil dari *clustering* didapatkan. Proses ini bertujuan untuk memvisualisasikan dan menginterpretasikan hasil *clustering* yang optimal dalam bentuk grafik dan tabel, serta mengetahui karakteristik dari masing-masing *cluster*. Selain itu proses visualisasi juga digunakan sebagai pertimbangan dalam tahapan evaluasi pada *cluster*.

E. Evaluasi hasil cluster

Pada tahapan ini dilakukan observasi hasil *clustering* pada tahap visualisasi. Hasil *clustering* dievaluasi jika *cluster* homogen dengan *cluster* lain. Artinya keragaman antar *cluster* dengan lainnya lebih besar dan keragaman didalam *cluster* lebih kecil. Hasil *clustering* dapat diukur dengan memperhatikan persentase distribusi data antar *cluster* pada atribut katagori *group overdue days*.



GAMBAR 2. Visualisasi piechart variabel group overdue days

Berdasarkan visualiasi hasil pengujian pengelompokan data menggunakan nilai k optimal 2, didapatkan 2 *cluster* dengan jumlah anggota *cluster 0* yaitu sebanyak 790 dan anggota *cluster 1* yaitu sebanyak 468. Pada hasil

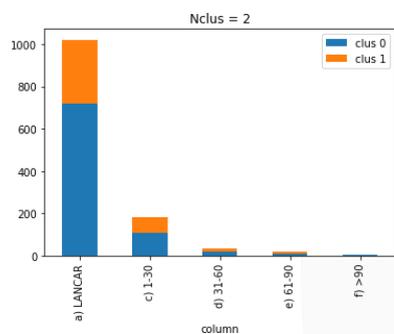
IV. HASIL DAN PEMBAHASAN

A. Hasil Clustering Algoritma K-prototype

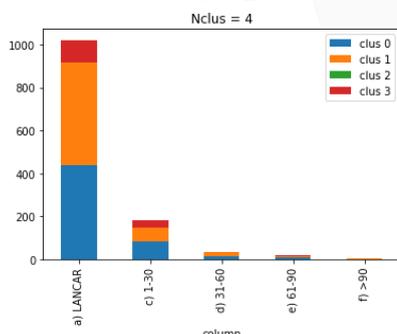
Implementasi algoritma *k-prototype* dilakukan pada data sampel berjumlah 1260 record. Penentuan jumlah *cluster* yang optimal dilakukan menginisiasi nilai k menggunakan perhitungan elbow dengan memasukan nilai $k = 1$ hingga $k = 10$.

visualisasi terlihat bahwa distribusi antar *cluster* menunjukkan proporsi data tidak terlalu signifikan untuk digunakan dalam menentukan *profiling* nasabah yang beresiko. Hal ini dikarenakan pada variabel kategori *grup overdue days* menunjukkan selisih proporsi antar *cluster* tidak terlalu besar yaitu selisih 1 % pada kategori lancar, sedangkan pada kategori lainnya juga menunjukkan hasil yang tidak jauh berbeda. Dari hasil pengujian diatas disimpulkan bahwa penentuan jumlah *cluster* optimal untuk digunakan dalam menentukan *profiling* resiko pada

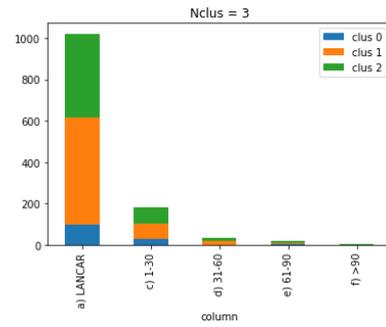
data *multifinance* belum bisa dilakukan. Untuk itu pada penelitian ini menyarankan untuk melakukan analisis lanjutan dalam menyelesaikan permasalahan analisis *profiling* resiko. Maka dari itu langkah yang diusulkan pada pengujian lanjutan yaitu dengan melakukan pengujian ulang pada tahap inisiasi nilai *k* dengan melakukan percobaan pada nilai *k* = 2 hingga *k* = 6. Implementasi pengujian jumlah *cluster* dilakukan atas dasar referensi penelitian sebelumnya dalam penentuan jumlah *cluster* optimal untuk segmentasi berdasarkan faktor pencari pada *cluster* [7].



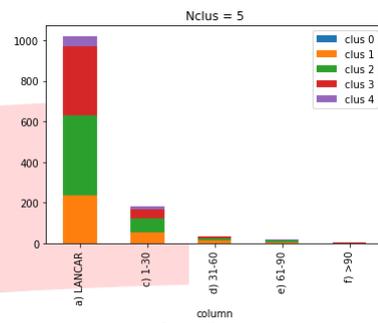
k = 2



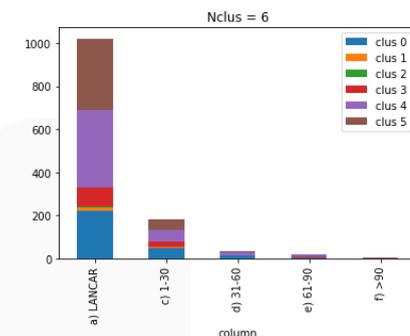
k = 4



k = 3

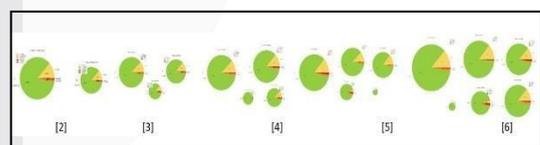


k = 5



k = 6

GAMBAR 3 visualisasi *barplot* pengujian nilai *k* = 2 hingga *k* = 6 kategori *group overdue days*



GAMBAR 4 visualisasi *pie chart* pengujian nilai *k* = 2 hingga *k* = 6 kategori *group overdue days*

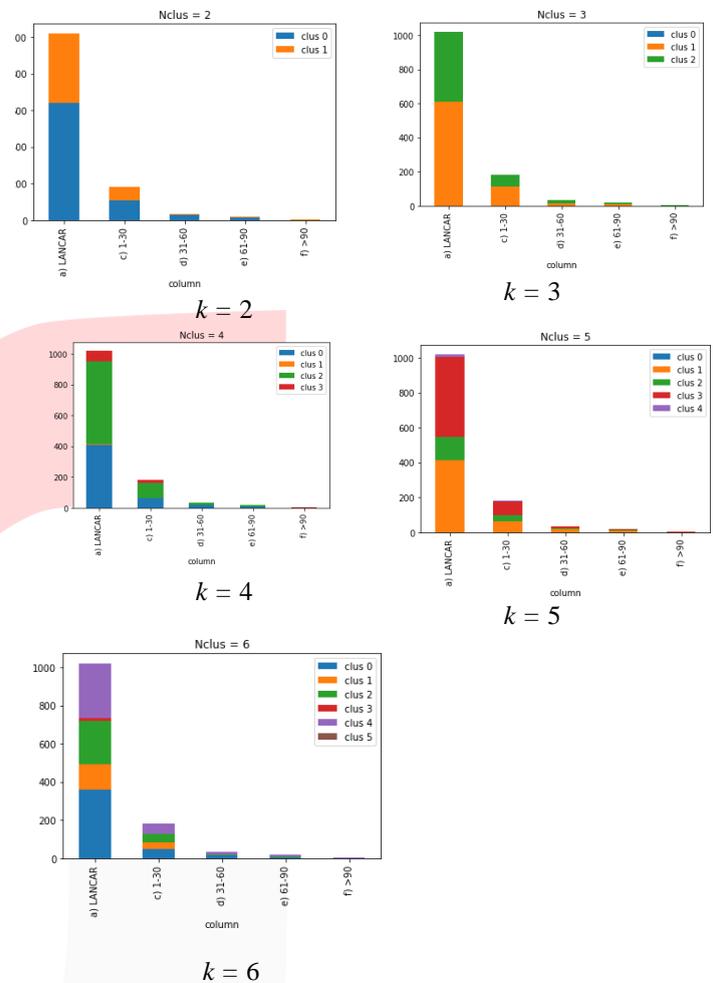
Gambar 3 menunjukkan hasil dari pengujian lanjutan dengan menggunakan nilai *k* = 2 hingga *k* = 6. Pada 6 hasil pengujian menunjukkan juga tidak menunjukkan hasil signifikan pada kategori *grup overdue days* yang beresiko dengan menggunakan dataset ini, namun ada sebuah perubahan *cluster* yang terbentuk dengan munculnya sebuah *cluster* baru dengan pola distribusi data yang berbeda. Pada pengujian nilai *k* = 5 dan nilai *k* = 6 ditemukan beberapa peningkatan *cluster* kategori lancar namun

ditemukan sebuah *anomaly* pada data yang menunjukkan sebuah *cluster* dengan nilai akurasi tinggi yaitu 100% dengan jumlah anggota 2 dengan karakteristik data sebagaiberikut :

Tabel 3. Variabel pada data nasabah dengan karakteristik lancar

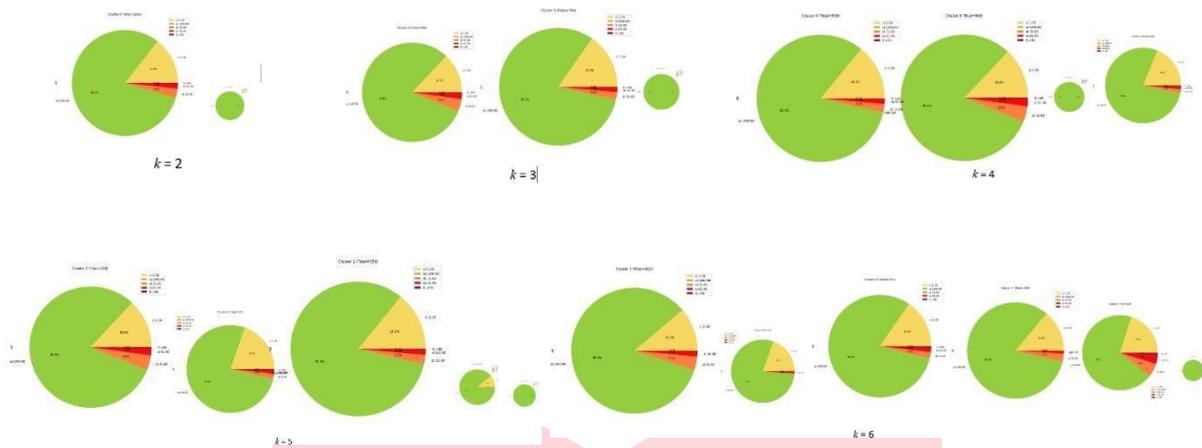
variabel		
	1	2
<i>Cust occu</i>	Wiraswasta	Pegawai swasta
<i>Salary</i>	3000000	4000000
<i>BPKB Name</i>	<i>Differrent</i>	<i>Differrent</i>
Harga motor	1320000	4600000
<i>Spouse salary</i>	0.0	0.0
Usia	39	22
Angsuran	564000	308000

Table 5 merupakan *cluster* dengan tingkat kelancaran 100%. Berdasarkan data pada tabel ditemukan bahwa beberapa variabel antar data tidak menunjukkan karakteristik yang kontras untuk mencirikan *profiling* resiko. Hal ini dikarenakan sistem kerja perusahaan *multifinance* melakukan *drop* data dengan nasabah dengan kategori keterlambatan lebih dari 90 hari dari daftar pantau marketing sehingga jumlah data nasabah dengan kategori keterlambatan cenderung sedikit ditemukan pada data sehingga menyebabkan hasil dari pengelompokan menunjukkan hasil rasio kelancaran lebih tinggi dibanding dengan nasabah beresiko. Selain itu terbatasnya akses penulis dalam mendapatkan data yang mendukung tentang *profiling* pada perusahaan *multifinance* terbatas, atas dasar pengujian diatas. Maka pengujian untuk menemukan *profiling* resiko dilakukan dengan menambahkan atribut rasio gaji dan angsuran untuk melihat apakah pola *clustering* berubah.



GAMBAR 5 visualisasi barplot pengujian nilai $k = 2$ hingga $k = 6$ kategori *group overdue days*

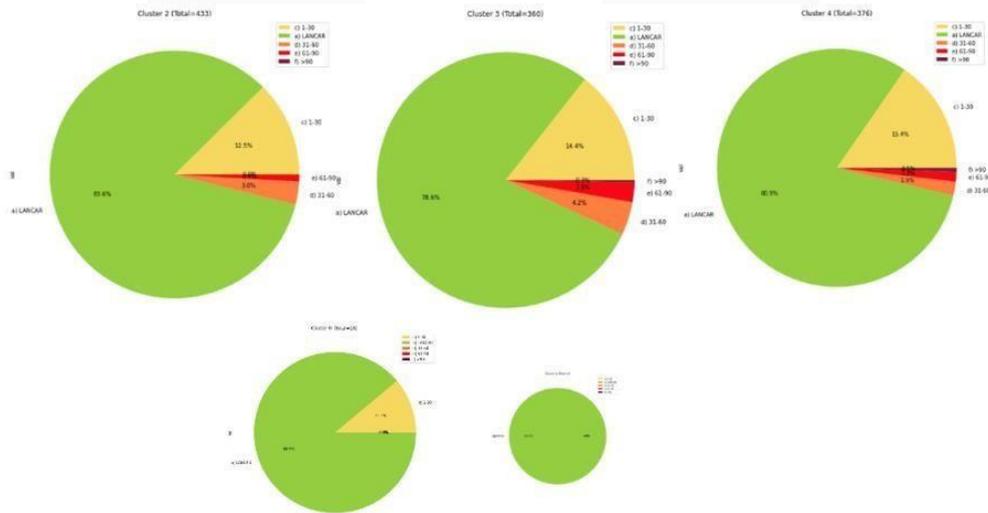
Gambar 12 menunjukkan hasil dari pengujian dengan menggunakan variabel *cust occu*, *salary*, angsuran, *group overdue days* dan 1 variabel tambahan yaitu perbandingan rasio *salary* terhadap angsuran dengan menggunakan nilai $k = 2$ hingga $k = 6$. Pada 6 hasil pengujian pada 4 variabel tersebut menunjukkan terbentuk sebuah pola *cluster* baru yang dapat dilihat pada visualisasi *pie chart* dibawah.



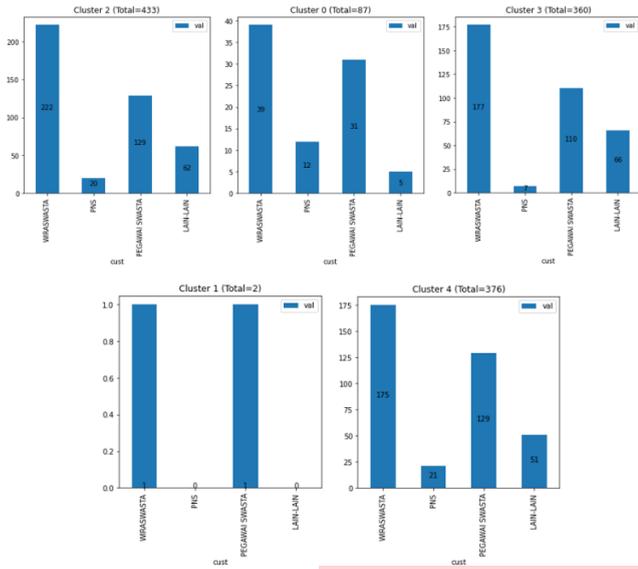
GAMBAR 6. visualisasi piechart pengujian nilai $k = 2$ hingga $k = 6$ kategori *group overdue days*

Pada gambar 13 visualisasi ukuran piechart merepresentasikan jumlah anggota pada setiap cluster terhadap setiap cluster yang terbentuk dengan 6 pengujian nilai $k = 2$ hingga $k = 6$. Pada variabel *group overdue days* menunjukkan bahwa distribusi cluster kelancaran yang hampir merata disetiap pengujian nilai k dengan selisih setiap pengujian tidak kurang dari 5%. Namun pada pengujian $k = 5$ ditemukan 2 cluster

dengan persentase keterlambatan paling rendah dibandingkan cluster lain yaitu 88.9% dengan keterlambatan 11% sedangkan untuk cluster setelahnya dengan presentasi kelancaran 100% dengan keterlambatan 0% maka dari penelitian ini melakukan pengujian pada nilai $k = 5$ untuk menganalisis tingkat kelancaran berdasarkan *profiling* resiko lancar.



Gambar 7. Visualisasi piechart variabel *group overdue days*



GAMBAR 8. Visualisasi *barplot* variabel *cust occu*

variabel	Rata rata keseluruhan data
Salary	3490063
Angsuran	528488
Rasio	16%

TABEL 5. Analisis *cluster* berdasarkan karakteristik nasabah lancar

Cluster	Karakteristik nasabah lancar
0	<ul style="list-style-type: none"> a. Rata rata pekerjaan nasabah pada <i>cluster</i> ini memiliki pekerjaan wiraswasta b. Rata rata jumlah gaji nasabah yang dimiliki diatas rata rata kisaran 4.7 juta c. Rata rata jumlah angsuran nasabah yang dimiliki diatas rata rata kisaran 620 ribu d. Rata rata rasio antara gaji dan angsuran nasabah yang dimiliki dibawah rata rata yaitu 13% e. Status keterlambatan group overdue days menunjukkan rasio kelancaran berkisaran 76.9%
1	<ul style="list-style-type: none"> a. Rata rata pekerjaan nasabah pada <i>cluster</i> ini memiliki pekerjaan wiraswasta b. Rata rata jumlah gaji nasabah yang dimiliki

	<ul style="list-style-type: none"> diatas rata rata kisaran 3.7 juta c. Rata rata jumlah angsuran nasabah yang dimiliki diatas rata rata kisaran 520 ribu d. Rata rata rasio antara gaji dan angsuran nasabah yang dimiliki dibawah rata rata yaitu 14% e. Status keterlambatan group overdue days menunjukkan rasio kelancaran berkisaran 82.3%
2	<ul style="list-style-type: none"> a. Rata rata pekerjaan nasabah pada <i>cluster</i> ini memiliki pekerjaan wiraswasta b. Rata rata jumlah gaji nasabah yang dimiliki dibawah rata rata kisaran 2.5 juta c. Rata rata jumlah angsuran nasabah yang dimiliki diatas rata rata kisaran 493 ribu

	<p>a. Rata rata rasio antara gaji dan angsuran nasabah yang dimiliki dibawah rata rata yaitu 20%</p> <p>b. Status keterlambatan group overdue days menunjukkan rasio kelancaran berkisaran 80.8%</p>
3	<p>a. Rata rata pekerjaan nasabah pada <i>cluster</i> ini memiliki pekerjaan pegawaiswasta</p> <p>b. Rata rata jumlah gaji nasabah yang dimiliki diatas rata rata kisaran 7.1 juta</p> <p>c. Rata rata jumlah angsuran nasabah yang dimiliki diatas rata rata kisaran 845 ribu</p> <p>d. Rata rata rasio antara gaji dan angsuran nasabah yang dimiliki dibawah rata rata yaitu 12%</p> <p>e. Status keterlambatan group overdue days menunjukkan rasio kelancaran berkisaran 88.9%</p>
4	<p>a. Rata rata pekerjaan nasabah pada <i>cluster</i> ini memiliki pekerjaan pegawaiswasta</p> <p>b. Rata rata jumlah gaji nasabah yang dimiliki dibawah rata rata kisaran 3.5 juta</p> <p>c. Rata rata jumlah angsuran nasabah yang dimiliki diatas rata rata kisaran 431 ribu</p> <p>d. Rata rata rasio antara gaji dan angsuran nasabah yang dimiliki dibawah rata rata yaitu 1.3%</p> <p>e. Status keterlambatan group overdue days menunjukkan rasio kelancaran berkisaran 100%</p>

V. KESIMPULAN

Berdasarkan hasil penelitian ini maka dapat disimpulkan:

- A. Penentuan *cluster* menggunakan metode elbow pada pengelompokan menggunakan algoritma *k-prototype* pada data *multifinance* belum dapat

digunakan untuk menyelesaikan permasalahan *profiling* resiko. karena proporsi antar *cluster* cenderung homogen artinya *cluster* yang terbentuk tidak memiliki karakteristik yang kuat untuk menentukan *profiling* nasabah yang beresiko.

- B. Algoritma *k-prototype* dapat digunakan dalam studi kasus *profiling* nasabah lancar dengan menganalisis karakteristik *cluster* yang terbentuk pada dataset dengan menambahkan atribut rasio.
- C. Hasil *clustering* pada penelitian ini menunjukkan kelompok dengan tingkat kelancaran tertinggi ada pada *cluster* 4 yaitu dengan tingkat kelancaran 100% dan jumlah rasio gaji terhadap angsuran nasabah dibawah rata rata yaitu 1.3% yang membuat *cluster* ini menjadi *cluster* dengan tingkat kelancaran tertinggi dibanding *cluster* lain

REFERENSI

- [1] SEMBIRING, JIMMY, HARRY J. SUMAMPOUW, and WILFRIED S. MANOPPO. "Analisis Kredit Bermasalah Pada PT. Adira Dinamika *Multifinance* Tbk Cabang Manado." JURNAL ADMINISTRASI BISNIS (JAB) 4.4 (2016).
- [2] Aritonang, Swandi. ANALISIS PENERAPAN SISTEM PENGENDALIAN INTERNAL PADA PROSES PEMBERIAN DANA PEMBIAYAAN DI PT. SINARMAS *MULTIFINANCE* CABANG CIKARANG. Diss. President University, 2017.
- [3] Yin, Shuang, et al. "Applications of *Clustering* with Mixed Type Data in Life Insurance." Risks 9.3 (2021):47.
- [4] Gan, G., Ma, C. and Wu, J., 2020. Data *clustering*: theory, algorithms, and applications. Society for Industrial and Applied Mathematics.
- [5] Huang, Zhexue. "*Clustering* large data sets with mixed numeric and categorical values." Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD). 1997.
- [6] Novidianto, Raditya, and Kartika Fithriasari. "Algoritma *ClusterMix K-Prototype* Untuk Menangkap Karakteristik Pasien Berdasarkan Variabel Penciri Mortalitas Pasien

- Dengan Gagal Jantung." *Inferensi* 4.1 (2021): 37-46.
- [7] Syakur, M. A., et al. "Integration k-means clustering method and elbow method for identification of the best customer profile cluster." *IOP conference series: materials science and engineering*. Vol. 336. No. 1. IOP Publishing, 2018
- [8] Sulthoni, Ahmad Shohibus, Rachmadita Andreswari, and Faqih Hamami. "Segmentasi Pelanggan Pt. Telekomunikasi Seluler Indonesia Menggunakan Clustering Algoritma K-prototype Dan Metode Elbow Sebagai Perumusan Strategi Marketing." *eProceedings of Engineering* 8.3 (2021).
- [9] Huang, Zhexue. "Extensions to the *k-means* algorithm for clustering large data sets with categorical values." *Data mining and knowledge discovery* 2.3 (1998): 283-304.
- [10] Huang, Zhexue. "Clustering large data sets with mixed numeric and categorical values." *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. 1997.
- [11] Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." *J* 2.2 (2019): 226-235.