SEGMENTASI PELANGGAN PT. TELEKOMUNIKASI SELULER INDONESIA MENGGUNAKAN CLUSTERING ALGORITMA K-PROTOTYPES DAN METODE ELBOW SEBAGAI PERUMUSAN STRATEGI MARKETING

CUSTOMER SEGMENTATION PT. TELEKOMUNIKASI SELULER INDONESIA USES CLUSTERING K-PROTOTYPES ALGORITHM AND ELBOW METHOD FOR FORMULATING MARKETING STRATEGY

Ahmad Shohibus Sulthoni[1], Rachmadita Andreswari [2], Faqih Hamami[3]

[1,2,3]Universitas Telkom, Bandung

[1]assulthoni@student.telkomuniversity.ac.id [2]tewidodo@telkomuniversity.ac.id
[3] faqihhamami@telkomuniversity.ac.id

Abstrak

PT. Telekomunikasi Selular Indonesia (Telkomsel) merupakan salah satu pemain besar di industri penyedia layanan seluler. Jumlah pelanggan Telkomsel pada tahun 2020 mencapai 163 Juta pelanggan aktif. Namun angka ini tidak mencerminkan kinerja pemasaran dari Telkomsel. Pada akhir 2019, posisi Telkomsel terancam oleh kompetitor seperti XL Axiata. Perusahaan XL Axiata mencatatkan penetrasi pelanggan baru lebih tinggi daripada Telkomsel.Salah satu cara untuk mempertahankan pangsa pasar di tengah ketatnya kompetisi adalah segmentasi pelanggan. Cara ini menghasilkan rekomendasi strategi yang dapat diterapkan oleh pihak pemasaran Telkomsel. Segmentasi pelanggan pada Telkomsel dilakukan dengan clustering dengan algoritma k-prototypes. Melalui metode elbow, penelitian ini mendapatkan jumlah cluster terbaik yaitu 4.Masing-masing cluster mewakili satu segmen dari pelanggan Telkomsel. Pada tiap segmen Telkomsel dapat menerapkan strategi yang berbeda. Untuk segmen pertama, rekomendasi strategi yang dihasilkan adalah menjalin kerja sama dengan merek OPPO. Sedangkan untuk segmen kedua strategi yang diterapkan adalah layanan khusus pengguna APPLE seperti iTunes. Untuk segmen ketiga, strategi yang dapat diterapkan adalah pengembangan produk digital BYU dan M2M. Segmen keempat strategi yang dapat diterapkan adalah kolaborasi bundling dengan merek XIAOMI.
Kata Kunci: segmentasi, clustering, k-prototypes, metode elbow

---

Abstract

PT. Telekomunikasi Selular Indonesia (Telkomsel) is one of the big players in the cellular service provider industry. The number of Telkomsel subscribers in 2020 reached 163 million active customers. However, this figure does not reflect the marketing performance of Telkomsel. At the end of 2019, Telkomsel's position was threatened by competitors such as XL Axiata. XL Axiata has recorded a higher penetration of new subscribers than Telkomsel.One way to maintain market share amidst intense competition is customer segmentation. This method produces strategic recommendations that can be applied by Telkomsel's marketing parties. Customer segmentation at Telkomsel is done by clustering with the k-prototypes algorithm. Through the elbow method, this study obtained the best number of clusters, which is 4.Each cluster represents one segment of Telkomsel's subscribers. In each segment, Telkomsel can apply different strategies. For the first segment, the resulting strategic recommendation is to collaborate with the OPPO brand. Meanwhile, for the second segment, the strategy implemented is special services for APPLE users such as iTunes. For the third segment, the strategy that can be implemented is the development of BYU and M2M digital products. The fourth segment of the strategy that can be implemented is the bundling collaboration with the XIAOMI brand..
Keyword : *segmentation, clustering, k-prototypes, elbow method*

---

1. Pendahuluan

Penetrasi internet di Indonesia pada saat ini cukup pesat. Berdasarkan survei Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), pada tahun 2018 terdapat 171,17 juta masyarakat Indonesia merupakan pengguna internet (APJII, 2019). Artinya terdapat 64,8 persen dari keseluruhan masyarakat Indonesia telah terkoneksi internet. Hal ini merupakan angka yang besar karena terjadi kenaikan 10,12 persen pengguna internet dari tahun 2017. Apabila dibandingkan dengan pertumbuhan penduduk yang rata-rata tumbuh 1,34 persen per tahun, persentase penetrasi

internet di Indonesia merupakan angka yang fantastis (Badan Pusat Statistik, 2018). Pertumbuhan angka pengguna internet juga diikuti dengan pertumbuhan penggunaan operator seluler. Sejalan dengan survei yang dilakukan APJII pada tahun 2018, sebanyak 96,6 persen pengguna terkoneksi dengan internet melalui paket data/kuota dari operator seluler di tahun 2019 (APJII, 2019). Dari beberapa daerah yang ada di Pulau Jawa, Provinsi Daerah Khusus Ibukota Jakarta (DKI Jakarta) memiliki persentase pengguna internet per jumlah penduduk yang sangat besar. DKI Jakarta memiliki penduduk 11.063.324 jiwa berdasarkan statistik pada tahun 2019 (BPS, 2019). Di samping itu, pengguna internet di DKI Jakarta pada tahun 2019 sebesar 8,9 Juta pengguna (Katadata, 2019). Artinya persentase pengguna internet di DKI Jakarta mencapai lebih dari 80 persen. Angka ini cukup besar dibandingkan dengan luas wilayah DKI Jakarta yang hanya 662,33 km2 (BPS, 2019).

PT. Telekomunikasi Seluler (Telkomsel) melayani masyarakat dengan berbagai jenis produk seluler yang ditawarkan. Beberapa jenis produk seluler yang ditawarkan adalah Simpati, Halo, ByU, AS dan LOOP. Setiap produk yang ditawarkan oleh Telkomsel memiliki peminat dan pelanggannya masing-masing. Di dalam setiap produk pengguna memiliki karakteristik dan kebiasaan yang berbeda-beda. Hal ini menjadi masalah baru ketika Telkomsel ingin memahami pelanggannya. Pemahaman pelanggan ini menjadi kunci dalam mengetahui preferensi pelanggan. Preferensi pelanggan sering kali dilakukan dengan cara segmentasi / pengelompokan (Zhao et al., 2010). Berdasarkan preferensi tersebut internal dari Telkomsel dapat melakukan berbagai hal antara lain dapat mengetahui posisi produk di pasar sehingga dapat melakukan penawaran produk secara efektif (Kashwan & Velu, 2013). Dengan karakteristik dan jumlah pelanggan yang sangat besar dan beragam, sulit dilakukan segmentasi secara manual. Secara tradisional, perusahaan yang bergerak di bidang telekomunikasi menggunakan atribut satu dimensi untuk segmentasi pelanggan yaitu kontribusi terhadap pendapatan perusahaan (Zhao et al., 2010).

Berdasarkan penelitian yang sudah ada serta permasalahan yang dialami oleh Telkomsel, maka terdapat urgensi untuk melakukan penelitian terkait segmentasi pelanggan di perusahaan Telkomsel khususnya wilayah DKI Jakarta. Pemilihan wilayah DKI Jakarta mempertimbangkan rasio pengguna internet per jumlah penduduk di wilayah tersebut yang besar. Dikarenakan belum ada penelitian yang spesifik membahas segmentasi pelanggan Telkomsel di DKI Jakarta dengan menggunakan data karakteristik pelanggan maka penelitian ini dirasa perlu untuk segera dilakukan. Selain itu, penelitian yang menggunakan metode elbow dan algoritma *K-means* dengan objek penelitian data telekomunikasi masih sedikit. Dengan sebab-sebab yang telah disebutkan, penulis merancang dan membuat penelitian yang berjudul "Segmentasi Pelanggan PT. Telekomunikasi Seluler Indonesia Menggunakan *Clustering* Algoritma *K-prototypes* Dan Metode *Elbow*". Alasan penggunaan algoritma *k-prototypes* adalah data yang digunakan penulis berbentuk tabel dengan nilai kategori dan numerik. Sehingga *k-prototypes* cocok untuk diimplementasikan pada data seperti ini. Selain itu, metode *elbow* juga termasuk metode yang tergolong populer untuk menemukan nilai parameter klaster yang tepat pada algoritma *k-prototype*.

2. Dasar Teori dan Metodologi

2.1 Dasar Teori

2.1.1 Data Mining

Data mining merupakan proses untuk menemukan pola dari sejumlah data. Sumber data bisa berupa basis data, pangkalan data, website maupun repositori lainnya (Jiawei et al., 2012). Data mining bukan merupakan disiplin ilmu yang baru. Pada era sebelumnya, data mining disebut dengan 'trawling', 'fishing through data' dan 'data dredging' (Hand, 2007).Hand menjelaskan bahwa data mining modern merupakan gabungan antara statistik, ilmu komputer, pembelajaran mesin dan teknologi basis data. Berdasarkan penjelasan tersebut dapat dilihat bahwa untuk melakukan proses-proses data mining membutuhkan beberapa disiplin ilmu.

2.1.2 Clustering

Clustering merupakan salah satu teknik data mining yang popular sebagai alat untuk mengelompokkan data berdasarkan kemiripan tanpa diketahui kelompok atau kelas sebelumnya. Pengelompokan didasarkan atas karakter data yang mirip satu sama lain. Clustering dikelompokkan dalam pembelajaran unsupervised. Karakteristik dari pembelajaran unsupervised antara lain tidak menerapkan pengawasan dari manusia melalui label atau target kelas (Xu & Tian, 2015).

2.1.3 K-Prototypes

K-prototypes digunakan untuk dataset yang memiliki tipe data campuran (numerik dan kategori). Algoritma *K-prototypes* mengintegrasikan *k-means* dengan *k-modes*. Algoritma *k-prototypes* secara praktis lebih berguna karena objek yang sering ditemui di basis data dunia nyata adalah objek tipe campuran (Huang, 1998). *K-prototypes* juga menggunakan ukuran ketidaksamaan antar objek pada dataset. Ukuran ketidaksamaan yang digunakan berbeda dengan *k-modes*. Dalam *k-prototypes* ukuran ketidaksamaan (*dissimilarity measure*) menggabungkan persamaan *euclidean distance* dengan *dissimilarity measure* yang ada dalam *k-modes*.

Secara matematis *dissimilarity measures* pada *k-prototypes* untuk objek X dan Y dengan tipe data campuran dapat didefinisikan sebagai berikut :

$$d_2(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

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

Dimana,

- d_2 merupakan simbol *dissimiliarity measures*
- γ merupakan bobot yang akan didapat setelah proses konvergen

2.1.4 Metode *Elbow*

Metode *elbow* hanya memberikan nilai k yang optimal. Optimasi nilai k didapatkan melalui fungsi biaya yang digunakan. Berikut fungsi biaya yang digunakan dalam *k-prototypes* :

$$P(W,Q) = \sum_{l=1}^{k}\left(\sum_{i=1}^{n} w_{i,j} \sum_{j=1}^{p}(x_{i,j} - q_{l,j})^2 + \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} \delta(x_{i,j}, q_{l,j})\right)$$

Hasil dari fungsi biaya ini disebut *inertia*. Semakin rendah nilai *inertia* menunjukkan bahwa jarak tiap titik data terhadap titik pusat klaster semakin rendah. Dalam pendekatan metode *elbow*, pemilihan nilai k optimal diketahui dari penurunan *inertia* yang curam sehingga membentuk sebuah siku (Yuan & Yang, 2019).

2.2 Sistematika Penelitian

2.2.1 Tahap Awal

Pada tahap awal dilakukan identifikasi masalah pada perusahaan Telkomsel di Provinsi DKI Jakarta. Kemudian melakukan studi literatur untuk menambah pengetahuan dan landasan teori atas solusi permasalahan. Setelah literatur terkumpul, langkah selanjutnya adalah observasi data dan lingkungan yang terkait. Selanjutnya menentukan rumusan masalah yang tepat sebagai dasar permasalahan penelitian. Terakhir adalah penentuan tujuan dari penelitian atas permasalahan yang telah dirumuskan.

2.2.2 Pengolahan Data

2.2.2.1 Pra-proses

Setelah data berhasil terkumpul, langkah selanjutnya adalah ekstraksi data. Ekstraksi yang dimaksud adalah proses pengubahan format penyimpanan data. Selanjutnya dilakukan seleksi data. Seleksi yang dimaksud adalah pemisahan data wilayah DKI Jakarta dengan wilayah lainnya. Selain itu, dilakukan pula seleksi atas fitur-fitur yang ada di dalam data. Fitur yang dimaksud adalah fitur yang masih berupa kolom pada data. Pemilihan fitur dilakukan berdasarkan kegunaan dan relevansi atas permasalahan yang telah dirumuskan. Setelah data terseleksi, langkah selanjutnya adalah transformasi data. Transformasi ditujukan untuk melakukan pengubahan wilayah dengan titik koordinat pusat wilayah tersebut. Wilayah yang digunakan adalah kecamatan. Penggunaan titik koordinat memudahkan dalam proses pemodelan dan visualisasi data.

2.2.2.2 Clustering Data

Pada bagian ini, langkah yang dilakukan adalah menentukan rentang jumlah klaster sebagai jumlah eksperimen. Rentang ini dibatasi agar metode elbow yang digunakan dapat dievaluasi dengan baik. Rentang dimulai dari angka 1 hingga n. Dengan menggunakan rentang ini, maka terdapat percobaan sebanyak n. Setiap percobaan akan dilakukan sub-proses clustering data. Sub-proses ini meliputi penentuan titik pusat sebagai centroid dari setiap klaster. Setelah itu dilakukan perhitungan jarak setiap titik data ke pusat klaster (centroid). Langkah selanjutnya adalah kalkulasi inertia. Sub-proses ini dilakukan berulang kali sebanyak n. Setelah perulangan dilakukan, selanjutnya penentuan jumlah klaster terbaik dengan menggunakan metode elbow. Setelah itu, dilakukan pengelompokan data menggunakan jumlah klaster yang optimal tersebut.

2.2.3 Tahap Akhir

Pada tahap akhir ini merupakan tahap untuk evaluasi hasil pengelompokan data. Evaluasi dilakukan mulai dari eksplorasi hasil pengelompokan. Eksplorasi ini dilakukan untuk mengetahui karakteristik setiap klaster. Setelah eksplorasi dilakukan, langkah selanjutnya adalah visualisasi hasil. Visualisasi ini sebagai cara mengkomunikasikan hasil penelitian kepada pihak terkait. Setelah itu, dilakukan pengambilan kesimpulan dan saran terhadap penelitian yang sudah dilakukan.

3. Analisis dan Perancangan

3.1 Pengumpulan Data

Dalam melakukan segmentasi pelanggan dengan clustering dibutuhkan pengumpulan data pelanggan Telkomsel. Data diperoleh dari perusahaan yang memiliki akses kepada basis data Telkomsel. Melalui skema kerja sama, penelitian ini

juga mendapatkan akses replikasi data yang dikirim melalui File Transfer Protocol (FTP). Server yang digunakan untuk menerima data adalah publicstorage01.telkomuniversity.ac.id.

3.2 Pra-pemrosesan Data

3.2.1 Ekstraksi dan Seleksi Data

Data yang berhasil terkumpul berbentuk *zip*. Data ini kemudian diekstraksi agar dapat dibaca secara mudah pada penelitian ini. Selain itu data juga dipilih berdasarkan lokasi DKI Jakarta. Kolom pada data mentah yang didapat terdiri dari 109 kolom. Nama-nama kolom tersebut antara lain:

[msisdn, lac, ci, hlrindex, mcc, mnc, ageoflocationinformation, imsi, skey, scpaddress, tvendor, ttype, imei, capedge, capgsmgprs, caphsdpa, caphsupa, capumts, spmms, sposname, spotap, lte, wlan, form, regional, propinsi, kabupaten, kecamatan, kelurahan, card_type, nam, ts11, ts12, ts21, ts22, ts61, ts62, bs26, bs30, plmnss5, camel_phase, odbic, odbroam, vlrnumber, sgsn_num, sgsn_address, subscriber_type, product_id, mscid, capgsmcsd, capgsmsms, frequency, osvendor, osversion, site_name, longitude, latitude, area, branch, sub_branch, cluster, address, node, cgi_prepaid, cgi_postpaid, home_reflex, zone1, zone2, zone3, zone4, zone5, scp_ip, scp_id, onlinestatus, latest_offline_time, pcrfsn, totalconsumption, sa, imeiflag, tqstartdate, tqenddate, timeusage, policyname, authmode, ocsitpl_name, tcsitpl_name, smscsitpl_name, ucsitpl_name, mcsitpl_name, vlrtpl_name, sgsntpl_name, sub_status, aaasn, eps, lteautoprov, mmehost, mmerealm, vplmn, papchap, simauth, peapauth, countryname, operatorname, domain, outroamer, homemcc, homemnc, homecountryname, homeoperatorname]

Dari keseluruhan kolom tidak semua kolom dapat digunakan dalam penelitian ini. Kolom-kolom yang digunakan merupakan kolom yang dapat mendukung penyelesaian masalah. Beberapa literatur menyebutkan bahwa solusi atas permasalahan segmentasi dapat diatasi dengan kolom demografi pelanggan, karakteristik pelanggan, kebiasaan pelanggan dan transaksi pelanggan (Hadi, 2019; Nisa et al., 2020; Zeniarja, 2015). Maka dari itu berdasarkan literatur dan deskripsi kolom yang ada, kolom yang digunakan untuk penelitian sebagai berikut.

Nama Kolom	Tipe Data	Deskripsi
Msisdn	Categorical	Kolom berisi nomor handphone pelanggan yang telah dilakukan penyamaran data dengan menggunakan kode samar "X"
lac_mask	Categorical	Kolom berisi kode area lokasi atau *Location Area Code* yang telah dilakukan penyamaran data dengan kode samar "X"
ci_mask	Categorical	Kolom berisi *Cell Identification* yang berisi nomor unik yang umumnya digunakan untuk mengidentifikasi setiap stasiun pemancar transceiver dasar atau sektor BTS dalam kode area lokasi jika tidak dalam jaringan GSM. Kolom ini juga dilakukan penyamaran data.
imsi	Categorical	Kolom berisi nomor *International Mobile Subscriber Identity* (IMSI). Setiap data merepresentasikan unique value dari setiap pelanggan. Nomor ini juga dilakukan penyamaran data.
imei	Categorical	Kolom berisikan nomor *International Mobile Equipment Identity*. Nomor ini digunakan sebagai penanda gawai yang terhubung dengan jaringan telekomunikasi. Setiap nomor merepresentasikan gawai yang berbeda. Nomor ini juga dilakukan penyamaran untuk privasi data.
skey	Number	Kolom berisi *Subscriber Key* yang merupakan nomor untuk mengidentifikasi kartu pelanggan berada pada rentang radio tertentu.
propinsi	Categorical	Kolom berisi nama propinsi pelanggan sesuai dengan aturan penamaan dari Badan Pusat Statistik.

Nama Kolom	Tipe Data	Deskripsi
kabupaten	Categorical	Kolom berisi nama kabupaten pelanggan sesuai dengan aturan penamaan dari Badan Pusat Statistik.
tvendor	Categorical	Kolom berisi nama vendor gawai yang digunakan oleh pelanggan seperti Samsung, Apple, Oppo dan yang lainnya.
sposname	Categorical	Kolom berisi nama Operating System (OS) yang digunakan oleh pelanggan seperti Android, IOS, Symbian dan yang lainnya.
lte	Categorical	Kolom berisi status *Long Term Evolution* (LTE) dari jaringan yang digunakan oleh pelanggan. Status bisa berupa YES atau NO.
subscriber_type	Categorical	Kolom berisi data tipe subsciber pelanggan. Tipe subscriber terbagi menjadi 3 yaitu : *Online Or Prepaid Subscriber, Hybrid Subscriber* dan Offline or Postpaid Subscriber.
domain	Categorical	Kolom berisi data jaringan yang digunakan oleh pelanggan. Data jaringan dapat berupa 3G atau 4G.
product_id	Categorical	Kolom berisi nama produk Telkomsel yang digunakan oleh pelanggan. Produk dari Telkomsel antara lain : Simpati, HALO Cek, AS, LOOP, HALO Hybrid, HALO VPN, Simpati Freedom, M2M, BYU.
totalconsumption	Number	Kolom berisi total bita yang telah dihabiskan pelanggan pada saat hari yang sama dengan proses pengambilan data yaitu 17 Juli 2020.

Kolom Msisdn, lac_mask, ci_mask, imsi, imei, skey digunakan sebagai indeks dari setiap baris. Sedangkan kolom selain beberapa kolom tersebut digunakan untuk proses clustering namum tetap harus melewati beberapa proses.

3.2.2 Pembersihan Data

Data yang berhasil terkumpul terdiri dari 4,5 juta baris. Namun ada banyak baris yang tidak lengkap atau mengandung *missing value*. Sehingga harus dilakukan proses pembersihan data. Persentase *missing value* dari tiap kolom sebagai berikut.

Nama Kolom	Persentase Missing Value (%)
tvendor	5.759228
sposname	9.941595
lte	5.773081
regional	0
propinsi	0
kabupaten	0
kecamatan	0
subscriber_type	0.000046
product_id	0.000046
totalconsumption	17.292612
domain	0.033133

Berdasarkan persentase tersebut dan hasil eksplorasi *missing value* yang ada di data merupakan ada beberapa *missing at random* dan *missing not at random*. Untuk kolom tvendor dan product_id, *missing value* yang ada pada data merupakan *missing at random*. Penanganan *missing value* seperti ini dapat dilakukan dengan cara drop baris dengan konsiderasi bahwa persentase juga sangat kecil. Namun untuk kolom totalconsumption berelasi dengan product_id, maka penanganan *missing value* dilakukan dengan cara mapping dengan product_id dengan pengisian rata-rata

konsumsi product_id yang ada (statisticsolutions.com, 2018). Sedangkan sposname berelasi kuat dengan tvendor, maka imputasi dilakukan dengan cara mapping dengan tvendor. *Missing value* yang tersisa kemudian dilakukan drop baris.

3.2.3 Transformasi Data

Setelah data bersih dari missing value selanjutnya dilakukan transformasi data. Transformasi dilakukan agar proses clustering berjalan dengan lancar dan sesuai dengan tujuan penelitian. Transformasi mencakup pengubahan lokasi pelanggan dengan titik koordinat kecamatan, transformasi nilai kategori yang mencakup pengubahan nilai yang jarang muncul untuk mereduksi nilai unik (unique value) dan transformasi skala numerik dengan Minimum-Maximum Scaling. Sampel data dapat dilihat pada table dibawah ini.

tvendor	sposname	lte	subscriber_type	product_id	domain	totalconsumption	x	y
13	0	1	2	7	1	0.00025	0.79986	0.15743
17	0	1	2	7	1	0.00015	0.62605	0.17175
19	5	0	1	6	0	0.0000	0.89865	0.20880
22	0	1	2	7	0	0.01630	0.68330	0.22423
12	3	0	0	2	0	0.04982	0.83580	0.00000

3.3 Implemenatasi dan Pengujian

3.3.1 Implementasi Algoritma K-Prototypes

Implementasi algoritma K-prototypes pada penelitian ini menggunakan pustaka (library) yang ada pada pemrograman python. Pustaka dibangun oleh Nico dan dipublikasikan pada https://pypi.org/project/kmodes/. Pustaka yang berlisensi MIT License ini terinspirasi dari penelitian yang dilakukan Huang pada 1998(Huang, 1998). Pustaka ini bersifat open-source yang berarti bahwa semua orang dapat berkontribusi dalam pengembangan pustaka ini dengan melalui prosedur tertentu.

Implementasi algoritma K-prototypes dilakukan pada data sampel berjumlah 9582 baris. Implementasi dilakukan dengan 13 kali percobaan dengan mengubah nilai k dari 2 hingga 15. Pemilihan rentang dilakukan atas dasar eksperimen dengan pertimbangan bahwa referensi terkait juga memakai rentang antara 2 hingga 20 (Saputra & Riksakomara, 2018; Syakur et al., 2018; Yuan & Yang, 2019)Berikut adalah langkah-langkah algoritma K-prototypes.

1. Penentuan nilai k
   Pada implementasi ini dilakukan percobaan nilai k = 2. Nilai k akan bertambah sesuai dengan percobaan yang dilakukan hingga k = 15. Pada masing-masing nilai k akan dilakukan komputasi dari langkah 2 hingga ke langkah 9. Lalu mengulangi seluruh dengan mengubah penambahan nilai k (increment).
2. Pemilihan centroid
   Centroid untuk iterasi awal dipilih secara acak namun untuk iterasi selanjutnya akan dipilih berdasarkan pembaruhan centroid. Pada tahap ini centroid 1 ditandai dengan baris berwarna hijau. Untuk centroid 2 ditandai dengan baris berwarna biru. Kedua centroid ini menjadi pusat masing-masing klaster hingga centroid diperbaharui sesuai fungsi biaya.
3. Perhitungan *dissimilarity measures* untuk fitur categorical
   Perhitungan dissimilarity measures untuk categorical dilakukan dengan ukuran matching dissimilarity. Ukuran ini mencocokkan fitur categorical tiap baris dengan fitur categorical tiap centroid. Dissimilarity measures pada fitur categorical dapat dihitung dengan menggunakan rumus $\delta(x_j, y_j)$. Rumus ini telah dijelaskan pada bagian sebelumnya.
4. Perhitungan *dissimiliarity measures* untuk fitur numerik
   Perhitungan *dissimilarity measures* pada fitur numerik dilakukan dengan cara menghitung euclidean distance dari masing-masing data. Euclidean distance yang dimaksud dapat dilihat pada persamaan berikut.
   $$distance(X, Y) = \sum_{j=1}^{p} (x_j - y_j)^2$$
5. Perhitungan fungsi biaya untuk tiap data dari setiap *cluster*
   Dengan menggunakan hasil dari matching dissimilarity dan euclidean distance dapat dilakukan kalkulasi fungsi biaya dari tiap data untuk tiap klaster. Fungsi biaya ini merupakan penjumlahan atas matching

dissimilarity dan euclidean distance yang dikalikan dengan gamma (\gamma). Perhitungan ini dapat direpresentasikan dengan menggunakan persamaan berikut.

$$d_2(X,Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j)$$

Dalam penelitian ini, gamma (γ) didapatkan dengan menggunakan estimasi yang telah didefinisikan oleh Huang (Huang, 1998). Estimasi dilakukan dengan persamaan berikut.

$$\gamma = 0.5 * std(X_{numerical})$$

$std(X_{numerical})$ merupakan standar deviasi dari fitur numeric yang terdapat dalam data. Pada sampel data yang digunakan dalam contoh perhitungan dalam excel, standar deviasi bernilai 0,1513.

6. Penetapan *cluster* pada setiap data berdasarkan fungsi biaya
   Total fungsi biaya didapatkan dengan menjumlahkan seluruh fungsi biaya dari tiap data yang telah diperoleh pada langkah 5. Penjumlahan fungsi biaya tersebut dapat dilakukan dengan persamaan berikut.
   $$\min(cost_{cluster_1}, \dots, cost_{cluster_n})$$
   Fungsi tersebut akan menghasilkan suatu nilai terkecil dari deretan nilai fungsi biaya (*array*). Kemudian ditentukan klaster dengan menggunakan indeks pada *array* tersebut.
7. Perhitungan total fungsi biaya
   $$\sum_{i=1}^{j} \left(\sum_{x=1}^{y} cost_{x,i} \right)$$
   Dimana,
   i = indeks untuk mengidentifikasi klaster ke-i
   j = batas indeks untuk mengidentifikasi nomor klaster terbesar
   x = indeks untuk mengidentifikasi baris dari tiap data
   y = indeks terakhir pada baris terakhir
   $cost_{x,i}$= fungsi biaya pada baris ke-x dan klaster ke-i
8. Pembaharuan centroid
   Pembaharuan centroid dilakukan dengan cara mencari modus dari setiap fitur categorical dan mencari rata-rata (means) dari fitur numerik. Perhitungan modus dan rata-rata dapat direpresentasikan dengan fungsi berikut.
   $$= MODE.SNGL(categorical_x)$$
   $$= AVERAGE(numeric_x)$$
   Fungsi ini diaplikasikan pada data tiap klaster. Sehingga centroid pada tiap klaster akan berubah.
9. Mengulangi langkah ke 2 hingga konvergen
   Setelah didapatkan centroid baru, selanjutnya adalah mengulangi langkah ke 2 hingga fungsi biaya tidak mengalami penurunan. Hal ini disebut dengan konvergen. Setelah kondisi konvergen tercapai maka menambahkan nilai k sebagai iterasi selanjutnya.

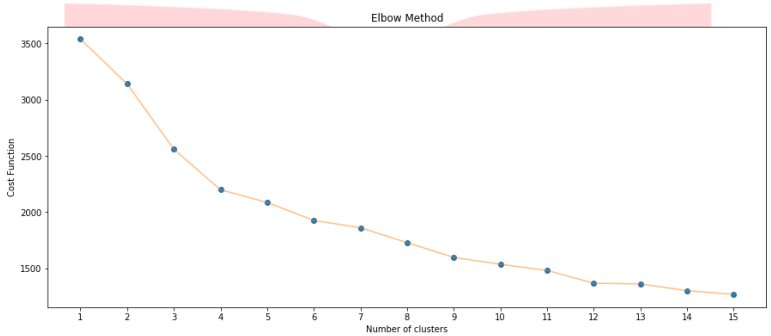
3.3.2 Pengujian Metode Elbow

Penerapan k-prototypes akan menghasilkan 15 nilai fungsi biaya. Dari 15 nilai ini akan diuji dengan menggunakan metode elbow sebagai dasar penentuan *cluster* terbaik. Berdasarkan eksperimen yang dilakukan maka nilai fungsi biaya dapat dilihat dari table berikut.

Jumlah klaster	Nilai Fungsi biaya	Perubahan
1	3538.2313696775996	N/A
2	3139.672735249663	-398.55863442793634
3	2558.656428611106	-581.0163066385571
4	2198.6323247845844	-360.0241038265217
5	2085.125620139946	-113.50670464463838
6	1925.5876100013595	-159.53801013858651

7	1860.4216396779825	-65.165970323377
8	1727.90039279055	-132.52124688743243
9	1595.8533318199147	-132.04706097063536
10	1534.9780416255367	-60.87529019437807
11	1480.2661379206102	-54.7119037049265
12	1365.9623970948799	-114.3037408257303
13	1359.3914428121927	-6.570954282687126
14	1300.0603799213957	-59.33106289079706
15	1267.0094019939104	-33.05097792748529

Nilai fungsi biaya ini kemudian divisualisasikan untuk mendapat titik siku (elbow) dari proses eksperimen yang dilakukan. Titik siku ini nantinya dapat dijadikan nilai k yang optimal. Hasil visualisasi dapat dilihat pada gambar dibawah ini.



Nilai siku yang dapat diambil dari eksperimen adalah 4. Sehingga penelitian ini menggunakan nilai 4 sebagai nilai k terbaik yang akan diimplementasi dalam proses *clustering* seluruh data. Namun, sebagai proses validasi maka diperlukan pengujian dengan silhouette score.

3.3.3 Pengujian Silhouette Score

Sebagai validasi hasil clustering dilakukan evaluasi dengan menggunakan silhouette score. Evaluasi ini dilakukan dengan bahasa pemrograman python pustaka scikit-learn. Potongan code untuk pengujian ini direpresentasikan sebagai berikut.

```
cluster = kproto.fit_predict(X_sample_np, categorical=[0,1,2,3,4,5])
    within_sum_squares.append(kproto.cost_)
    if i == 1:
        continue
score = silhouette_score(X_sample_np[:,:6], cluster, metric='jaccard')
```

Potongan code tersebut diimplementasikan di setiap iterasi percobaan. Sehingga akan menghasilkan 15 nilai silhouette score. Pada perhitungan silhouette ini digunakan metric jaccard dikarenakan jaccard dapat menangani tipe data kategori dengan baik (Hwang et al., 2018). Secara matematis jaccard dapat diformulasikan dengan persamaan berikut (Hwang et al., 2018).

$$jac(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

Nilai yang dihasilkan dari silhouette score ini berada pada rentang 0 hingga 1. Semakin tinggi nilai, maka hasil eksperimen semakin baik. Hasil dari seluruh silhouette score dapat dilihat pada table berikut.

Iterasi	Score
1	N/A
2	0.27813928567442636
3	0.32132812045552117
4	0.41715905812173815
5	0.23446303231940686
6	0.15565003641672717
7	0.18068929369031791

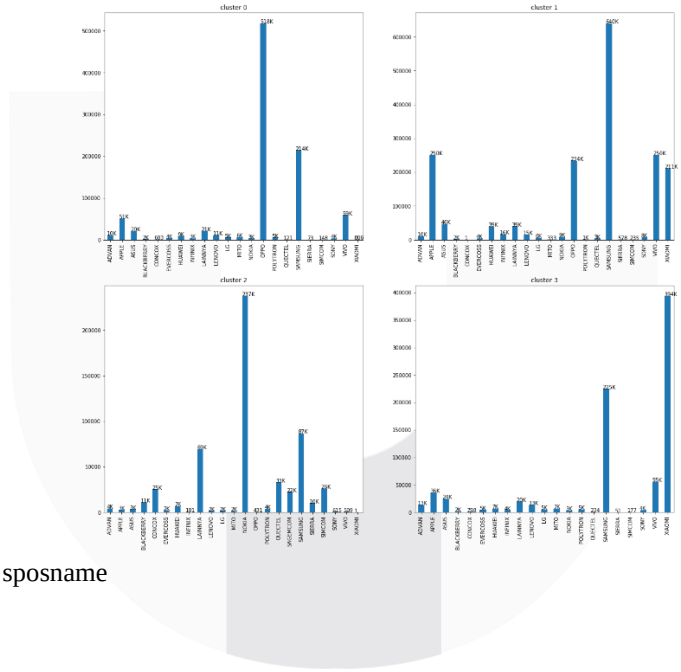
Iterasi	Score
8	0.16784725768052985
9	0.21266932006010714
10	0.21167832103424275
11	0.06279498139600541
12	0.04769940397274211
13	0.10802329345602528
14	0.10740227693101313
15	0.011795741747311848

3.3.4 Interpretasi Cluster

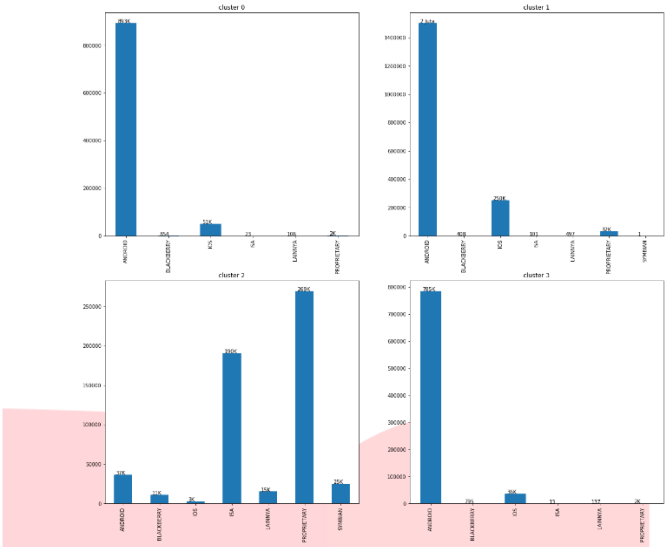
Setelah dilakukan proses *clustering* langkah selanjutnya adalah melakukan interpretasi atas cluster yang terbentuk. Proses ini dilakukan agar mendapatkan solusi atas permasalahan yang sedang dialami. Interpretasi cluster dilakukan dengan cara memvisualisasikan hasil clustering dan menarik kesimpulan atas visualisasi tersebut. Jumlah baris dari setiap cluster dapat dilihat pada tabel dibawah.

Klaster	Jumlah Baris
0	947132
1	1785827
2	550959
3	823130

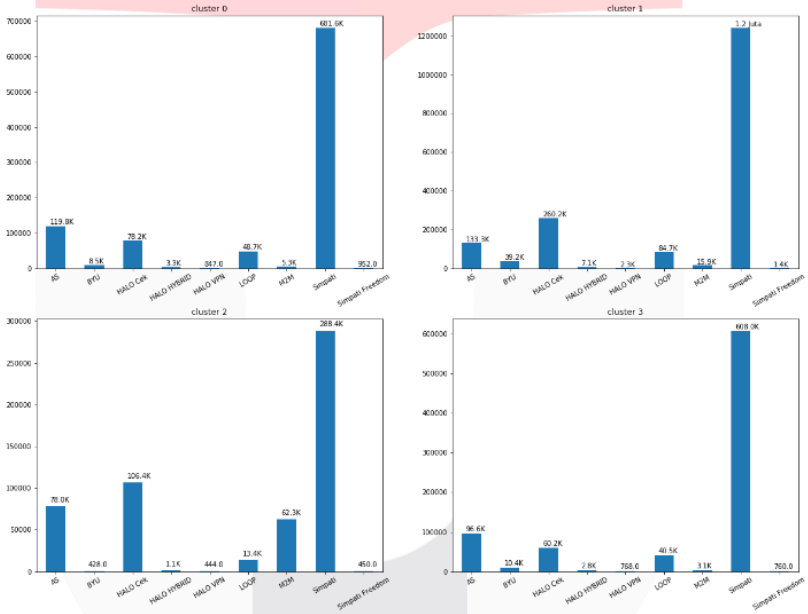
a) Visualisasi Kolom tvendor



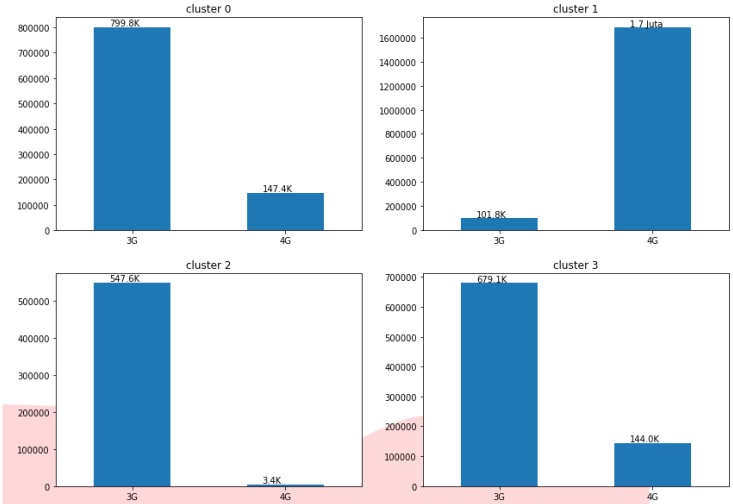
b) Visualisasi kolom sposname



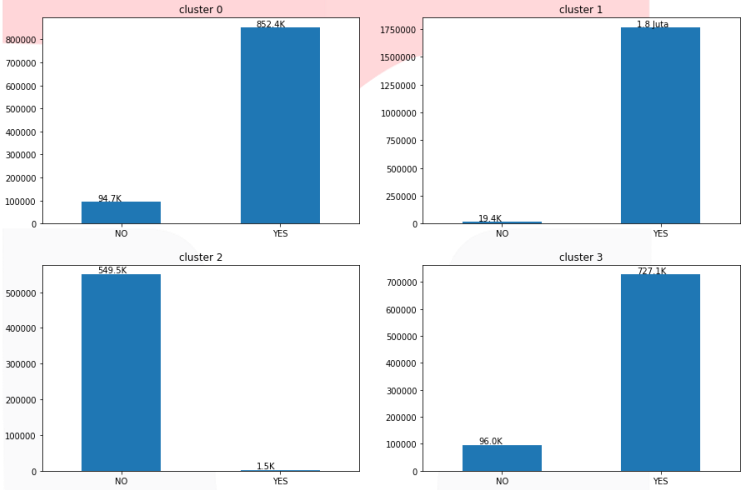
c) Visualisasi kolom product_id



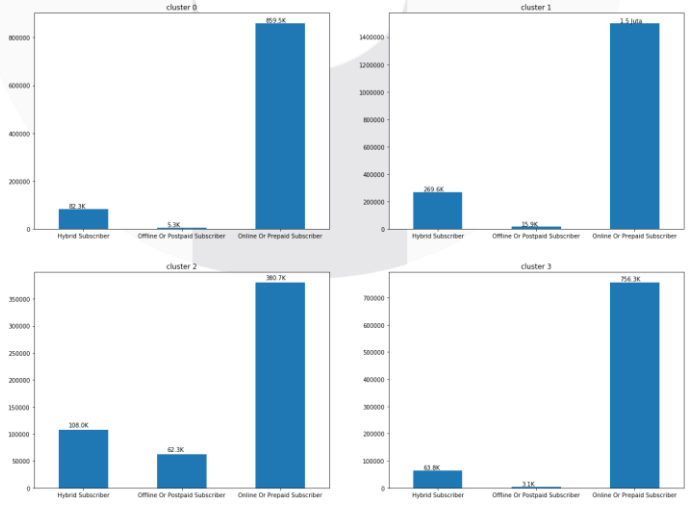
d) Visualisasi kolom domain



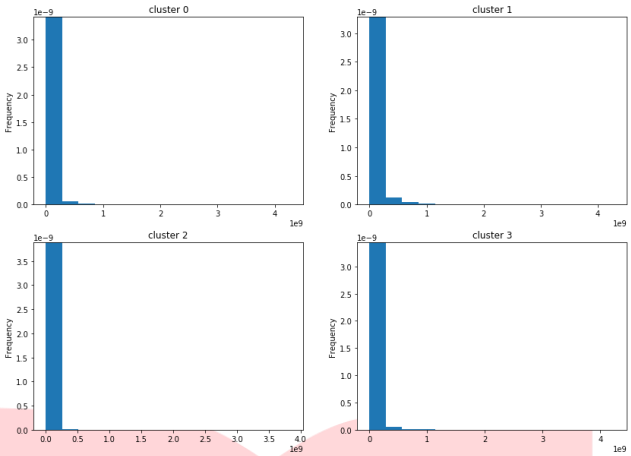
e) Visualisasi kolom lte



f) Visualisasi kolom subscriber_type

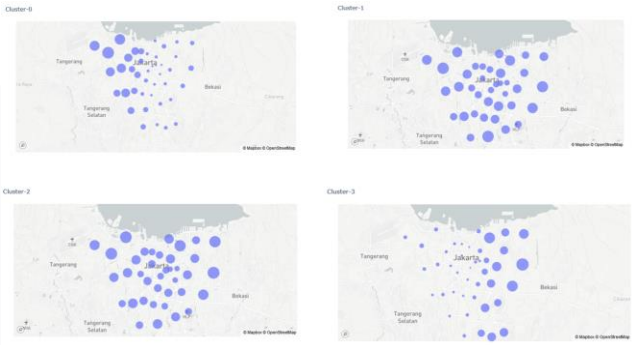


g) Visualisasi kolom totalconsumption



cluster	count	mean	std
0	947.132	29.222.772,4	126.885.933,5
1	1.785.827	58.129.304,5	195.402.152,7
2	550.959	31.631.956	67.048.700,7
3	823.130	27.488.355,4	124.799.757

h) Visualisasi lokasi (x dan y)



i) Penarikan kesimpulan strategi

Cluster	Karakteristik	Rekomendasi Strategi
0	Vendor OPPO dan SAMSUNG mendominasi secara jumlah. Mayoritas pelanggan di cluster ini menggunakan produk SIMPATI dengan domain 3G dan status LTE aktif. Pelanggan pada cluster ini adalah pelanggan jenis prabayar (*online or prepaid subscriber*). Konsumsi data yang dilakukan oleh pelanggan memiliki rata-rata 29 *Megabytes* per hari.	Menjalin kerja sama dengan vendor gawai untuk memberikan promosi yang dapat meningkatkan loyalitas pelanggan. Memastikan jaringan 3G masih dapat menikmati layanan Telkomsel secara keseluruhan. Memberikan promosi yang dapat meningkatkan konsumsi data per hari. Fokus pemasaran dapat dilakukan di daerah barat jakarta
1	Pelanggan pada cluster ini mayoritas adalah pengguna OPPO. Namun mayoritas pengguna APPLE secara keseluruhan berada pada cluster ini. Jumlah pengguna XIAOMI dan VIVO juga cukup banyak pada cluster ini. Produk yang banyak digunakan pada cluster ini adalah SIMPATI dan Halo Cek dengan domain jaringan 4G dan status LTE aktif. Tipe pelanggan didominasi pelanggan prabayar dan hybrid. Rata-rata konsumsi data adalah 58 *Megabytes* per hari.	Memperhatikan pengguna APPLE dengan memberikan tawaran menarik. Menjalin kerja sama dengan OPPO selaku vendor yang memiliki pengguna terbanyak pada cluster ini. Memastikan jenis pelanggan Hybrid dapat menikmati layanan yang menarik dari Telkomsel. Fokus pemasaran dapat dilakukan di daerah pusat jakarta
2	Cluster ini banyak terdapat pengguna dengan vendor yang jarang ditemui di Indonesia seperti QUECTEL, NOKIA dan SIMCOM. Produk yang banyak digunakan adalah SIMPATI dan M2M dengan domain 3G dan status LTE tidak aktif. Pelanggan pada cluster ini merupakan pelanggan prabayar dengan konsumsi data 31 *Megabytes* per hari.	Memastikan bahwa pelanggan dengan merek gawai yang jarang tetap bisa mendapatkan kualitas layanan terbaik. Mendorong penggunaan 4G pada cluster ini. Memberikan penawaran istimewa untuk produk M2M & BYU sebagai produk digital yang dimiliki Telkomsel. Fokus pemasaran dapat dilakukan di daerah pusat jakarta
3	Vendor gawai XIAOMI sangat mendominasi secara jumlah. Pelanggan pada cluster ini mayoritas adalah pelanggan SIMPATI dan AS dengan domain jaringan 3G dan status LTE aktif. Rata-rata konsumsi data pada cluster ini adalah 27 *Megabytes* per hari	Kolaborasi dengan XIAOMI untuk memberikan *bundling* pada produk baru XIAOMI. Memberikan promosi yang ekstra kepada pelanggan di cluster ini agar dapat meningkatkan rata-rata konsumsi data per hari. Fokus pemasaran dapat dilakukan di daerah timur jakarta

4. Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan hasil penelitian ini, maka dapat disimpulkan:

1. Algoritma *K-prototypes* dapat digunakan dalam studi kasus segmentasi pelanggan Telkomsel. Algoritma dapat mengatasi fitur numerik dan kategori dengan baik.
2. Metode Elbow dapat digunakan pada algoritma *K-prototypes* sebagai dasar penentuan parameter jumlah klaster yang optimal.

3. Setiap klaster yang dihasilkan dari proses clustering masih saling tumpang tindih antar fitur. Namun setiap klaster memiliki karakteristik unik yang dapat dieksplorasi. Eksplorasi menghasilkan interpretasi strategi sebagai berikut. Untuk segmen pertama, rekomendasi strategi yang dihasilkan adalah menjalin kerja sama dengan merek OPPO. Sedangkan untuk segmen kedua strategi yang diterapkan adalah layanan khusus pengguna APPLE seperti *iTunes*. Untuk segmen ketiga, strategi yang dapat diterapkan adalah pengembangan produk digital BYU dan M2M. Segmen keempat strategi yang dapat diterapkan adalah kolaborasi *bundling* dengan merek XIAOMI.

4.2 Saran

Penelitian ini dapat disempurnakan dengan menambah atribut lain sehingga klaster yang dihasilkan akan lebih baik. Atribut yang dimaksud dapat diperoleh dari sumber data eksternal seperti harga gawai, lama berlangganan dan perilaku digital dari setiap pelanggan. Hal ini dapat menambah informasi yang didapatkan dari proses segmentasi.

Referensi

Hadi, A. (2019). Segmentasi Pelanggan Internet Service Provider (ISP) Berbasis Pillar K-Means. *Jurnal Ilmiah Teknologi Informasi Asia*, *13*(2), 151. https://doi.org/10.32815/jitika.v13i2.413

Hand, D. J. (2007). Principles of data mining. In *Drug Safety* (Vol. 30, Issue 7). https://doi.org/10.2165/00002018-200730070-00010

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery 2, 283-304. *Data Mining and Knowledge Discovery*, *2*(3), 283–304. https://www.researchgate.net/publication/220451944_Huang_Z_Extensions_to_the_k-Means_Algorithm_for_Clustering_Large_Data_Sets_with_Categorical_Values_Data_Mining_and_Knowledge_Discovery_2_283-304

Hwang, C.-M., Yang, M.-S., & Hung, W.-L. (2018). New similarity measures of intuitionistic fuzzy sets based on the Jaccard index with its application to clustering. *International Journal of Intelligent Systems*, *33*(8), 1672–1688. https://doi.org/10.1002/int.21990

Jiawei, H., Jian, P., & Micheline, K. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann. https://books.google.co.id/books?id=pQws07tdpjoC&lpg=PP1&ots=tzLt3Tnz2-&dq=data mining&lr&pg=PR6#v=onepage&q=data mining&f=false

Nisa, H. H., Cahyo, P., & Fikri, A. N. (2020). View of Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python. *Jurnal Manajemen Informatika*. https://search.unikom.ac.id/index.php/jamika/article/view/2683/1874

Saputra, D. B., & Riksakomara, E. (2018). Implementasi Fuzzy C-Means dan Model RFM untuk Segmentasi Pelanggan (Studi Kasus : PT. XYZ). *Jurnal Teknik ITS*, *7*(1). https://doi.org/10.12962/j23373539.v7i1.28230

statisticsolutions.com. (2018). *Missing Values in Data - Statistics Solutions*. https://www.statisticssolutions.com/missing-values-in-data/

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering*, *336*(1). https://doi.org/10.1088/1757-899X/336/1/012017

Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, *2*(2), 165–193. https://doi.org/10.1007/s40745-015-0040-1

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

Zeniarja, J. (2015). Prediksi Churn dan Segmentasi Pelanggan Menggunakan Backpropagation Neural Network Berbasis Evolution Strategies. *Publikasi.Dinus.Ac.Id*. http://publikasi.dinus.ac.id/index.php/technoc/article/view/706