Jurnal JAMIKA Manajemen Informatika

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Segmentasi Pelanggan Produk *Digital Service* Indihome Menggunakan Algoritma K-Means Berbasis Python

Nisa Hanum Harani[1], **Cahyo Prianto[2]**, **Fikri Aldi Nugraha[3]**
Program Studi Teknik Informatika, Politeknik Pos Indonesia, Bandung, Indonesia[1,2,3]
e-mail: nisahanum@poltekpos.ac.id[1], cahyoprianto@poltekpos.ac.id[2],
fikrialdinugraha@gmail.com[3]

Abstrak

PT. Telekomunikasi Indonesia adalah salah satu perusahaan yang mengedepankan pelanggan akan tetapi belum ada informasi tentang karakteristik pelanggan. Pada penelitian ini dilakukan analisa karakteristik pelanggan sebagai dasar penetapan segmentasi pelanggan dan customer profiling pelanggan produk digital service add on Indihome menggunakan Algoritma K-Means. Penentuan jumlah cluster terbaik dilakukan menggunakan metode Elbow dan diperoleh nilai K = 3, sehingga data pelanggan dikelompokkan kedalam tiga segmen. Pengolahan data pelanggan dibagi menjadi 3 simulasi dengan persentase data train dan data test 80% - 20%, 70% - 30% dan 50% - 50%. Data yang digunakan berjumlah 1392 record sebagai populasi dimana data tersebut akan digunakan untuk mencari karakteristik setiap data tersebut. Evaluasi cluster dilakukan menggunakan metode Silhouette Index, Davies Bouldin Index dan Calinski Harabasz Index. Hasil dari penelitian menunjukan bahwa simulasi ketiga merupakan simulasi terbaik berdasarka evaluasi cluster dengan presentasi data train 50% dan data test 50% dimana customer profiling dilihat dengan menganalisis anggota masing-masing cluster dari simulasi ketiga dimana cluster 0 memiliki anggota 396 pelanggan dengan kategori pelanggan yang memberikan keuntungan terbesar bagi perusahaan, cluster 1 memiliki anggota 286 pelanggan dengan kategori pelanggan yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan, dan cluster 2 memiliki anggota 14 pelanggan dengan kategori pelanggan yang memberikan keuntungan lebih sedikit daripada biaya untuk memberikan pelayanan.

***Kata Kunci** : K-Means, Segmentasi Pelanggan, Cluster, Customer Profiling*

Abstract

PT. Telekomunikasi Indonesia is one of the companies that prioritize customers, but there is no information about customer characteristics. In this research, an analysis of customer characteristics used as a basis for determining customer segmentation and customer profiling for digital products add on Indihome services using the K-Means Algorithm. Determination of the best number of clusters done using the Elbow method and a value of K = 3 obtained, so that customer data grouped into three segments. Customer data processing is divided into 3 simulations with the percentage of train data and test data 80% - 20%, 70% - 30% and 50% - 50%. The data used totaled 1392 records as a population where the data will used to find the characteristics of each data. Cluster evaluations carried out using the Silhouette Index, Davies Bouldin Index, and Calinski Harabasz Index methods. The results of the study show that the third simulation is the best based on cluster evaluation with 50% data train presentation and 50% data test where customer profiling is seen by analyzing the members of each cluster from the third simulation where cluster 0 has 396 customer members with a customer category that provides the biggest profit for the company, cluster 1 has members of 286 customers in the category of customers who unwittingly have great potential in providing benefits for the company, and cluster 2 has a member of 14 customers in the customer category that provides fewer benefits than the cost of providing services.

***Keywords** : K-Means, Customer Segmentation, Cluster, Customer Profiling*

Jurnal JAMIKA Manajemen Informatika

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

1. Pendahuluan

Dalam bisnis apapun baik dalam bidang jasa maupun manufaktur salah satu faktor yang mempengaruhi kemajuan sebuah perusahaan adalah pemasaran. Pemasaran bukan saja hanya pengembangan produk dan jasa yang dibutuhkan tetapi segmentasi pelanggan juga harus diperhitungkan [1]. PT. Telekomunikasi Indonesia melakukan kegiatan strategi pemasaran dengan cara publisitas produk pada sosial media, *personal selling*, serta periklanan. Namun pada kegiatan pemasaran tersebut dirasa belum efektif dikarenakan masih ada beberapa produk yang kurang diminati. Fokus utama perusahaan untuk bersaing dengan kompetitornya adalah pelanggan. Permasalahannya adalah belum ada informasi tentang karakteristik pelanggan, maka pada penelitian ini dilakukannya analisa karakteristik pelanggan sebagai dasar penetapan segmentasi pelanggan dan *customer profiling [2]*. Tujuan dari proses segmentasi pelanggan adalah untuk mengetahui perilaku pelanggan dan menerapkan strategi pemasaran yang tepat sehingga mendatangkan keuntungan bagi pihak perusahaan. Proses *marketing* (komunikasi, produk/jasa, program) dapat menjadi lebih terfokus karena masing-masing segmen memang sudah memiliki kemiripan, baik dari segi kebutuhan maupun perilakunya [3]. Pengelompokkan tersebut dilakukan dengan menggunakan teknik *data mining*. *Clustering* memiliki peran yang penting dalam *data mining*, dimana teknik ini akan membagi data kedalam beberapa *cluster* sesuai dengan kemiripannya [4]. K-Means merupakan salah satu algoritma data *clustering* non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih *cluster*/kelompok [5]. Adapun proses pengelompokkan data dilakukan dengan mengambil 3 jenis atribut diantaranya yaitu lama berlangganan, jumlah paket yang diambil, dan jumlah tagihan yang dibayar oleh pelanggan. Hasil dari pemodelan data tersebut akan mengelompokkan pelanggan kedalam sejumlah *cluster* dan menentukan *customer profiling*. Pengelompokkan tersebut akan menghasilkan karakteristik dari masing-masing pelanggan pada tiap *cluster* yang dapat dijadikan salah satu acuan untuk pengambilan keputusan perusahaan.

2. Kajian Pustaka

2.1 Segmentasi Pelanggan

Segmentasi terus menjadi konsep pemasaran yang penting juga dalam konteks *relationship marketing*. Meningkatkan hubungan dengan pelanggan menjadi lebih menarik dan akan menghasilkan pemahaman yang lebih baik tentang kebutuhan pelanggan. Segmentasi adalah proses membagi pelanggan menjadi beberapa *cluster* dengan kategori loyalitas pelanggan untuk membangun strategi pemasaran. Segmentasi pelanggan adalah salah satu langkah awal dalam membuat model bisnis [6].

2.2 Customer Profiling

Customer Profiling merupakan langkah yang yang dilakukan untuk memetakan dan mendalami profil pelanggan dengan lebih baik. Pemetaan profil konsumen dapat dilakukan dengan kombinasi data eksplisit (informasi mengenai konsumen yang didapatkan dari proses pendaftaran dan survei) dan data implisit (informasi perilaku konsumen yang didapatkan dengan pengamatan langsung) [7].

2.3 Algoritma K-Means

K-means merupakan metode *clustering* secara *partitioning* yang memisahkan data ke dalam kelompok yang berbeda. Dengan *partitioning* secara iteratif, K-Means mampu meminimalkan rata-rata jarak setiap data ke *cluster*-nya [8].

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Terdapat beberapa ukuran jarak yang digunakan sebagai ukuran kemiripan suatu instance data, salah satunya adalah jarak Euclidean. Perhitungan jarak Euclidean seperti pada persamaan 1 [9].

$$d(Xi, Cj) = \sqrt{\sum_{i=1}^{N} (Xi - Cj)^2} \quad (1)$$

Duran dan Odell (1974) menyatakan jika semakin kecil, kesamaan antara dua *d*(*Xi*,Cj) unit pengamatan semakin dekat. Syarat menggunakan jarak Euclid adalah jika semua fitur dalam dataset tidak saling berkorelasi. Jika terdapat fitur yang berkorelasi maka menggunakan konsep jarak Mahalanobis. Agusta (2007) menyatakan kelanjutan dari jarak tersebut dicari yang terdekat sehingga data akan mengelompok berdasarkan centroid yang paling dekat. Tahap berikutnya adalah update titik centroid dengan menghitung rata-rata jarak seluruh data terhadap centroid. Selanjutnya akan kembali lagi ke proses awal. Iterasi ini akan diulangi terus sampai didapatkan centroid yang konstan artinya titik centroid sudah tidak berubah lagi. Atau iterasi dihentikan berdasarkan jumlah iterasi maksimal yang ditentukan [9].

2.4 Python

Python adalah bahasa pemrograman yang bersifat *open source*. Bahasa pemrograman ini dioptimalisasikan untuk *software quality*, *developer productivity*, program *portability*, dan *component integration* (Lutz, 2010). Python telah digunakan untuk mengembangkan berbagai macam perangkat lunak, seperti *internet scripting*, *systems programming*, *user interfaces*, *product customization*, *numeric programming* dll. Python saat ini telah menduduki posisi empat atau lima bahasa pemrograman paling sering digunakan di seluruh dunia (Lutz, 2010) [10].

2.5 Metode Penentuan K (Metode Elbow)

Metode Elbow merupakan suatu metode yang dapat digunakan untuk menghasilkan informasi dalam menentukan jumlah *cluster* terbaik dengan cara melihat persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk siku pada suatu titik tertentu. Metode ini memberikan ide/gagasan dengan cara memilih nilai *cluster* dan kemudian menambah nilai *cluster* tersebut untuk dijadikan model data dalam penentuan *cluster* terbaik. Selain itu, persentase perhitungan yang dihasilkan menjadi pembanding antara jumlah *cluster* yang ditambah. Hasil persentase yang berbeda dari setiap nilai *cluster* dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya. Jika nilai *cluster* pertama dengan nilai *cluster* kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka nilai *cluster* tersebut yang terbaik [11].

2.6 Metode Evaluasi Cluster

2.6.1 Silhouette Index

Secara umum, indeks validitas *Silhouette* menghitung rata-rata nilai setiap titik pada himpunan data. Lebih spesifik, perhitungan nilai setiap titik adalah selisih nilai *separation* dan *compactness* yang dibagi dengan maksimum antara keduanya. Jumlah klaster yang terbaik ditunjukkan dengan nilai Silhouette yang semakin mendekati 1 (Rosseeuw, 1987) [12].

2.6.2 Davies-Bouldin Index

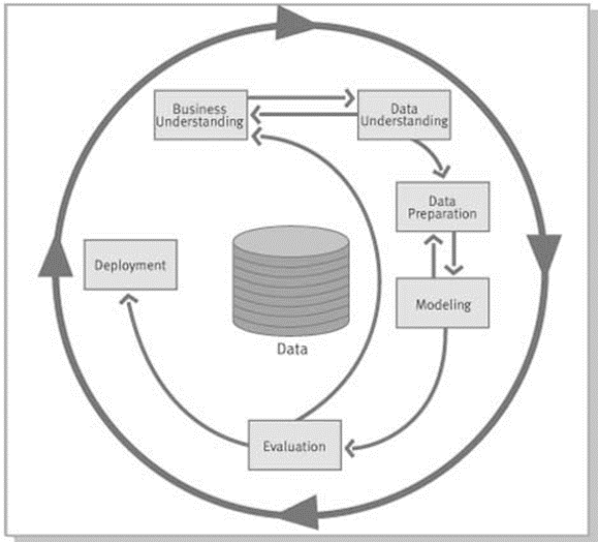
Indeks validitas Davies-Bouldin (DB) menghitung rata-rata nilai setiap titik pada himpunan data. Perhitungan nilai setiap titik adalah jumlah nilai *compactness* yang dibagi dengan jarak antara kedua titik pusat klaster sebagai *separation*. Jumlah *cluster* terbaik ditunjukkan dengan nilai DB yang semakin kecil (Davies & Bouldin, 1979) [12].

2.6.3 Calinski Harabasz Index

Indeks validitas Calinski-Harabasz (CH) menghitung perbandingan antara nilai *Sum of Square between cluster* (SSB) sebagai *separation* dan nilai *Sum of Square within cluster* (SSW) sebagai *compactness* yang dikalikan dengan faktor normalisasi, yaitu selisih jumlah data dengan jumlah *cluster* dibagi dengan jumlah *cluster* dikurang satu. Jumlah *cluster* terbaik ditunjukkan dengan semakin besar nilai CH (Baarsch & Celebi, 2012) [12].

3. Metode Penelitian

Metodologi penelitian yang digunakan adalah CRISP-DM (*Cross Standart Industries for Data Mining*). Metodologi CRISP-DM adalah standarisasi data mining yang disusun oleh tiga penggagas data mining market yaitu Daimcler Chrysler (Daimler Benz), SPSS, NCR. Kemudian dikembangkan pada berbagai *workshop* (antara tahun 1997-1999) [4].



Gambar 1. CRISP-DM

3.1 Tahapan-Tahapan Diagram Alur Penelitian

Terdapat beberapa tahapan dalam CRISP-DM (*Cross Industries Process for Data Mining*), namun pada penelitian ini tahapan yang dilakukan hanya sampai pada tahapan *evaluation* dikarenakan tujuan dari penelitian ini adalah untuk mengetahui data pelanggan terbagi ke dalam berapa segmen serta untuk memperoleh pengetahuan terkait dengan profil pelanggan berdasarkan segmentasinya [13].

3.1.1 Business Understanding

Fase ini meliputi penentuan tujuan bisnis, menilai situasi saat ini, menetapkan tujuan *data mining*, dan mengembangkan rencana penelitian. Tujuan bisnis yang akan

dilakukan pada penelitian ini adalah bagaimana melakukan segmentasi data pelanggan dan *customer profiling* menggunakan metode K-Means.

3.1.2 Data Understanding

Data yang digunakan dalam penelitian berjumlah 1392 *record* sebagai populasi dimana data tersebut diambil dari periode bulan januari sampai dengan bulan oktober berdasarkan berkas yang ditunjukan oleh objek penelitian, jika data yang diperoleh semakin banyak maka hasil dari akurasi datanya juga akan maksimal. Jika diperlukan acuan pustaka juga dapat dilakukan pada tahapan ini.

3.1.3 Data Preparation

Banyak persiapan yang dilakukan pada tahapan ini sehingga tak jarang fase ini juga disebut sebagai fase padat karya. Beberapa kegiatan seperti pemilihan tabel dan *field* terjadi pada fase ini. Pemilihan tabel dan *field* tersebut akan dimasukan atau ditransformasikan kedalam database yang lain atau *database* baru sebagai bahan atau *data mining* mentah. Atribut yang digunakan sebagai bahan *data mining* dibagi menjadi 3 kelompok atau 3 bagian yaitu data lama berlangganan, jumlah layanan yang diambil dan total tagihan pelanggan.

3.1.4 Modeling

Pada fase pemodelan dilakukan dengan menggunakan aplikasi jupyter notebook, dan dimasukkan juga metode K-Means. Dari data atribut yang telah dipilih pada fase *data preparation* yaitu data lama berlangganan, jumlah layanan yang diambil dan total tagihan pelanggan digunakan sebagai parameter untuk melakukan *clustering*.

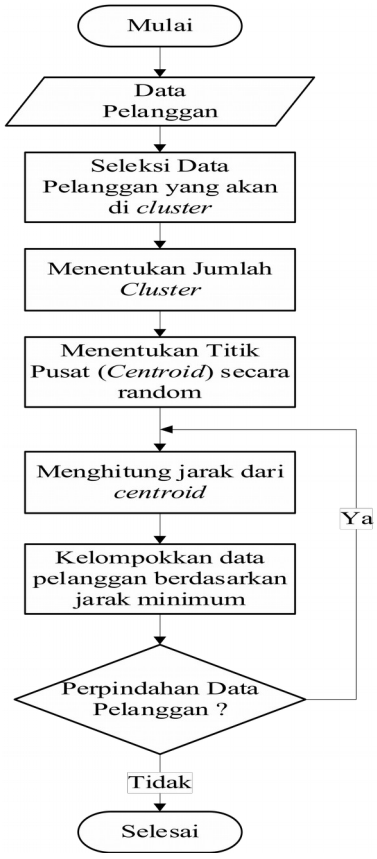
3.1.5 Evaluation

Fase ini merupakan tahapan analisa yang merupakan hasil dari pengolahan fase sebelumnya dengan menginterpretasikan data yang kemudian diperoleh segmentasi pelanggan dan *customer profiling*.

3.2 Penerapan Algoritma K-Means

K-Means merupakan salah satu algoritma dalam data mining yang bisa digunakan untuk melakukan pengelompokan/*clustering* suatu data [14]. Adapun diagram alir segmentasi pelanggan dengan menggunakan algoritma K-Means seperti pada gambar 2.

Alur segmentasi pelanggan dimulai dari menyiapkan data pelanggan, menyeleksi data pelanggan yang akan di*cluster,* menentukan jumlah *cluster*, menentukan titik pusat (*centroid*) secara random, menghitung jarak dari *centroid*, mengelompokkan data pelanggan berdasarkan jarak minimum, dan perpindahan data pelanggan. Jika terjadi perpindahan data pelanggan maka akan kembali ke proses menghitung jarak dari centroid. Jika tidak terjadi perpindahan data pelanggan maka proses selesai.



Gambar 2. Diagram Alir Segmentasi Data Pelanggan Dengan Algoritma K-Means

4. Hasil dan Pembahasan

4.1 Pengolahan Data Menggunakan Algoritma K-Means

Data yang digunakan dalam penelitian ini merupakan data pelanggan yang terdiri dari tiga jenis data yaitu data pelanggan indihome, data pelanggan *add on* serta data *churn* pelanggan. Adapun keseluruhan data pelanggan berjumlah 4640 *record,* namun pada penelitian ini data yang digunakan berjumlah 1392 record atau 30% dari jumlah keseluruhan data. Berikut ini merupakan penjelasan dari masing-masing data.

4.1.1 Data Pelanggan

1. Data Pelanggan Indihome
   Data pelanggan indihome merupakan data pelanggan yang berlangganan layanan jaringan internet. Pada data pelanggan, terdapat atribut seperti NCLI, ND_INTERNET, ND, CITEM_SPEEDY, KECEPATAN, DESKRIPSI, TGL_REG, TGL_ETAT, NAMA.
2. Data pelanggan *Add On*
   Merupakan data pelanggan yang berlangganan layanan tambahan produk digital yang disediakan oleh perusahaan untuk melengkapi layanan internet indihome. Data pelanggan *add on* memiliki atribut diantaranya WITEL, NCLI, NDOS, NDEM, NO_INET, ITEM, PRICE, TGL_VA, TGL_PS, KCONTACT.
3. Data *Churn* Pelanggan
   Data *Churn* pelanggan merupakan data pelanggan yang berhenti berlangganan layanan. Pada data ini terdapat beberapa atribut diantaranya adalah KAWASAN, WITEL, DATEL, NCLI, ND_INTERNET, DESKRIPSI, TGL_REG, TGL_ETAT, serta STATUS_ORDER.

Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

4.1.2 Seleksi Data Pelanggan

Tahap ini merupakan tahap persiapan data (*data preparation*), dimana proses yang dilakukan pada tahap ini yaitu pemilihan atribut yang akan digunakan selama proses *clustering*. Atribut yang terpilih merupakan atribut yang dapat mewakili identitas setiap pelanggan diantaranya adalah nomor pelanggan atau NCLI, lama berlangganan, jumlah layanan yang digunakan serta total tagihan. Namun dalam proses *clustering,* atribut yang digunakan hanya 3 atribut kecuali nomor pelanggan atau NCLI.

4.1.3 Eksperimen

Proses yang dilakukan pada tahap ini merupakan pemodelan, dimana tujuan dari pemodelan adalah untuk menganalisa dan memberi prediksi yang dapat mendekati kenyataan sebelum sistem di terapkan di lapangan [15]. Pemodelan dilakukan dengan menggunakan bahasa pemrograman python serta dibuat dalam 3 simulasi.

4.1.3.1 Import Module

Tahap ini merupakan proses import module yang digunakan dalam pemodelan.

```
In [1]: %matplotlib inline
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import numpy as np
        import sklearn
        from sklearn.model_selection import train_test_split
        from sklearn.cluster import KMeans
        from sklearn import metrics
        from sklearn.preprocessing import MinMaxScaler
        from sklearn.metrics import silhouette_score
        from sklearn.metrics import davies_bouldin_score
        from sklearn.metrics import pairwise_distances
```

Gambar 3. *Import Module*

4.1.3.2 Baca Data

Pada tahap ini merupakan proses memuat data (*load data*) yang akan diproses seperti pada gambar 4.

```
In [2]: data_cust = pd.read_csv('DATASET_MENTAH.csv')

        data_cust.head()
```

Out[2]:

	NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	JUMLAH_TAGIHAN_HARUS_DIBAYAR
0	39684298	11	1	418000
1	39713960	11	2	511500
2	39716635	11	2	698500
3	39818227	11	2	291500
4	68228	11	2	517000

Gambar 4. Baca Data

4.1.3.3 Normalisasi Data

Proses normalisasi data dilakukan merubah nilai data atau untuk menyamakan skala atribut data kedalam *range* yang lebih spesifik yang lebih kecil yaitu antara 0 – 1 [16]. Pada penelitian ini standar yang digunakan untuk melakukan normalisasi yaitu

MinMaxScaller. Untuk melihat perbedaan data pada sebelum dan sesudah proses normalisasi dapat dilihat pada Tabel 1 dan Tabel 2.

Tabel 1. Data Sebelum Normalisasi

NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	JUMLAH_TAGIHAN_HARUS_DIBAYAR
39684298	11	1	418000
39713960	11	2	511500
39716635	11	2	698500
39818227	11	2	291500
68228	11	2	517000

Tabel 2. Data Setelah Normalisasi

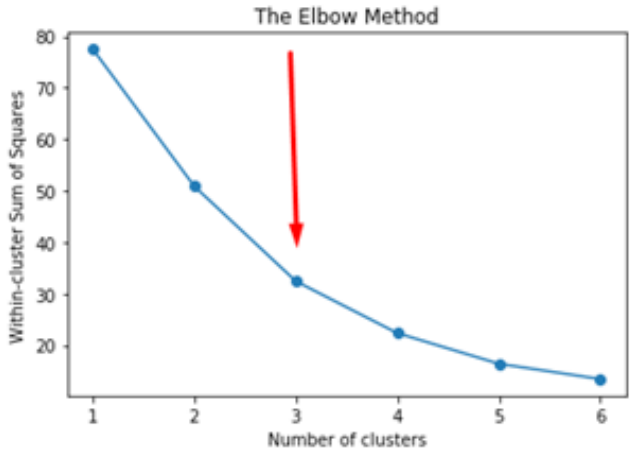
NCLI	LAMA_LANGGANAN	JUMLAH_LAYANAN	JUMLAH_TAGIHAN_HARUS_DIBAYAR
39684298	1.00	0.00	0.08
39713960	1.00	0.25	0.13
39716635	1.00	0.25	0.23
39818227	1.00	0.25	0.01
68228	1.00	0.25	0.13

4.1.3.4 *Split Data*

Split Data dilakukan untuk membagi data ke dalam *data train* dan *data test*. Dimana simulasi 1 data dibagi menjadi 80% *data train* atau 1113 *record* dan 20% *data test* atau 279 *record*, simulasi 2 data dibagi menjadi 70% *data train* atau 974 *record* dan 30% *data test* 418 *record*, kemudian simulasi 3 data dibagi menjadi 50% *data train* atau 696 *record* dan 50% *data test* atau 696 *record*.

4.1.3.5 *Penentuan Jumlah Cluster Terbaik*

Pada tahap ini penentuan cluster terbaik dilakukan dengan menggunakan metode Elbow dengan cara menghitung nilai *Sum Square Error* (SSE) kemudian memvisualisasikannya dalam bentuk grafik dimana hasil perhitungan SSE digambarkan dalam bentuk siku lalu nilai yang memiliki penuruan secara drastis merupakan jumlah K yang optimal yang ditunjukan pada Gambar 5 [11].



Gambar 5. Penentuan Jumlah Cluster Terbaik

4.1.3.6 *K-Means Clustering*

a. Simulasi 1

Pada simulasi 1 peneliti menggunakan persentasi pembagian *data train* dan *data test* (80% dan 20%). Proses segmentasi pada simulasi 1 dimulai dengan melakukan inisialisasi data yang bertujuan untuk mendeklarasikan atau menentukan data mana yang akan digunakan sebagai *data train* dan *data test* serta atribut yang digunakan untuk proses *clustering* yaitu lama langganan, jumlah layanan, dan total tagihan, dengan pendeklarasian sebagai berikut X = train_data2[['lama_langganan', 'jumlah_layanan', 'total_tagihan']]. Kemudian proses selanjutnya adalah pembuatan model *clustering* untuk simulasi 1 dengan jumlah K = 3 sesuai dengan hasil dari penentuan jumlah *cluster* terbaik dengan menggunakan metode Elbow dengan deklarasi sebagai berikut km = KMeans(n_clusters=3) dan ditambahkan perintah km.fit(X). Setelah pembuatan model, proses berikutnya adalah menentukan centroid atau titik pusat yang digunakan untuk melakukan proses selanjutnya yaitu *clustering* menggunakan algoritma K-Means dengan perintah centroid_train_data = km.cluster_centers_ dimana nilai dari centroid tersebut ditentukan secara *random* dan untuk melakukan *clustering* digunakan perintah km.predict(X). Setelah proses *clustering* selesai, dilakukan evaluasi *cluster* dengan menggunakan metode Silhouette Index (SI), Davies-Bouldin Index, dan Calinski Harabasz Index (CHI) yang mana hasil dari proses evaluasi tersebut dapat dilihat pada Tabel 3.

b. Simulasi 2

Pada simulasi 2 peneliti menggunakan persentasi pembagian *data train* dan *data test* 70% 30%. Untuk proses inisialisasi data, model *clustering*, penentuan centroid, dan *clustering* data pada simulasi kedua memiliki cara yang sama dengan simulasi 1, yang membedakan hanya jumlah *data train* dan *data test* sesuai dengan data yang telah dibagi pada proses *split data* simulasi kedua. Kemudian hasil dari evaluasi *cluster* yang dilakukan pada simulasi 2 dapat dilihat pada Tabel 4.

c. Simulasi 3

Pada simulasi peneliti menggunakan persentasi pembagian *data train* dan *data test* 50% 50%. Untuk proses inisialisasi data, model *clustering*, penentuan centroid, dan *clustering* data pada simulasi 3 memiliki cara yang sama dengan simulasi 1 dan simulasi 2, dimana yang membedakan simulasi 3 dengan simulasi yang lain hanya jumlah *data train* dan *data test* sesuai dengan data yang telah dibagi pada proses *split data* simulasi ketiga dan hasil evaluasi *cluster* yang dilakukan pada simulasi 3 dapat dilihat pada Tabel 5.

4.2 Pengkajian

Pada tahap ini, peneliti akan mengkaji terkait dengan penelitian yang dilakukan. Tujuan dilakukannya penelitian adalah untuk melakukan segmentasi data pelanggan serta melakukan *customer profiling* dengan menggunakan algoritma K-Means, dimana data pelanggan diperoleh dari arsip perusahaan.

Proses *clustering* yang dilakukan menggunakan simulasi pertama membagi data menjadi 80% *data training* atau 1113 *record* dan *data testing* 20% atau 279 *record* ke dalam 3 *cluster*. Hasil dari proses *clustering* menggunakan *data train* membagi data ke dalam *cluster* 0 berjumlah 479 *record*, *cluster* 1 berjumlah 607 *record* dan *cluster* 2 berjumlah 27 *record*. Sedangkan *data test* terbagi ke dalam *cluster* 0 sebanyak 163 *record*, *cluster* 1 sebanyak 108 *record*, dan *cluster* 2 sebanyak 8 *record*.

Selanjutnya proses *clustering* yang dilakukan menggunakan simulasi kedua membagi data menjadi 70% *data training* atau 974 *record* dan *data testing* 30% atau 418

record, dimana hasil *clustering* menggunakan *data training* membagi data ke dalam *cluster* 0 sebanyak 532 *record*, *cluster* 1 sebanyak 417 *record*, dan *cluster* 2 sebanyak 25 *record*. Kemudian hasil *clustering* menggunakan *data testing* membagi data ke dalam *cluster* 0 sebanyak 202 *record*, *cluster* 1 sebanyak 208 *record*¸ *cluster* 2 sebanyak 8 *record*.

Pada simulasi ketiga, data pelanggan dibagi menjadi 50% *data training* atau 696 *record* dan *data testing* 50% atau 696 *record*. Kemudian hasil *clustering* dengan menggunakan *data training* membagi data ke dalam *cluster* 0 sebanyak 396 *record*, *cluster* 1 sebanyak 286 *record*, dan *cluster* 2 sebanyak 14 *record*. Selain itu hasil *clustering* menggunakan *data testing* membagi data kedalam *cluster* 0 sebanyak 374 *record*, *cluster* 1 sebanyak 301 *record*, dan *cluster* 2 sebanyak 21 *record*.

Hasil *clustering* yang telah terbentuk dievaluasi *performance* nya untuk menentukan simulasi mana yang paling tepat digunakan untuk *clustering* data pelanggan tersebut. Evaluasi dilakukan dengan menghitung *score* dari Silhouette Index (SI), Davies-Bouldin Index, dan Calinski Harabasz Index (CHI) dimana hasil dari evaluasi tersebut dapat dilihat pada Tabel 3, Tabel 4 dan Tabel 5.

Tabel 3. Hasil Evaluasi *Performance* K-Means Simulasi 1

CLUSTER	SIMULASI 1					
	DATA TRAIN (80%)			DATA TEST (20%)		
	SI	DBI	CHI	SI	DBI	CHI
Global	0.535	0.670	806.107	0.508	0.630	198.522
0	0.482	0.848	598.288	0.470	1.053	114.979
1	0.500	0.964	567.455	0.433	0.896	113.259
2	0.703	0.603	360.728	0.717	0.426	117.956

Pada Tabel 3 simulasi 1 terdapat *data train* (80%) dan *data test* (20%) untuk mengetahui simulasi terbaik maka dilakukan validasi menggunakan indeks SI, DBI, dan CHI. Validasi tersebut dilakukan pada keseluruhan *data train* dan *data test* dengan hasil perhitungan diberi label Global. Kemudian, validasi juga dilakukan pada masing-masing *cluster* yang mana hasil perhitungannya diberi label 0, 1, 2 sesuai dengan *cluster.* Untuk mempermudah membandingkan hasil dari setiap simulasi, maka dipilihlah hasil validasi dengan label Global, dimana nilai validasi dari masing-masing indeks pada *data train* dan *data test* dihitung nilai rata-ratanya sehingga menghasilkan nilai sebagai berikut 0.521, 0.650, 502.315.

Tabel 4. Hasil Evaluasi Performance K-Means Simulasi 2

CLUSTER	SIMULASI 2					
	DATA TRAIN (70%)			DATA TEST (30%)		
	SI	DBI	CHI	SI	DBI	CHI
Global	0.539	0.641	751.767	0.508	0.731	251.635
0	0.474	0.853	474.070	0.470	0.880	236.459
1	0.498	0.988	451.299	0.484	0.977	231.581
2	0.716	0.543	384.504	0.658	0.677	93.711

Pada Tabel 4 simulasi 2 terdapat *data train* (70%) dan *data test* (30%) untuk mengetahui simulasi terbaik maka dilakukan validasi menggunakan indeks SI, DBI, dan CHI. Validasi tersebut dilakukan pada keseluruhan *data train* dan *data test* dengan hasil perhitungan diberi label Global. Kemudian, validasi juga dilakukan pada masing-masing *cluster* yang mana hasil perhitungannya diberi label 0, 1, 2 sesuai dengan *cluster.* Untuk mempermudah membandingkan hasil dari setiap simulasi, maka dipilihlah hasil validasi dengan label Global, dimana nilai validasi dari masing-masing indeks pada *data train* dan *data test* dihitung nilai rata-ratanya sehingga menghasilkan nilai sebagai berikut 0.524, 0.686, 501.701.

Jurnal JAMIKA Manajemen Informatika
Jurnal Manajemen Informatika (JAMIKA)
Volume 10 Nomor 2 Edisi Oktober 2020
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Tabel 5. Hasil Evaluasi *Performance* K-Means Simulasi 3

CLUSTER	SIMULASI 3					
	DATA TRAIN (50%)			DATA TEST (50%)		
	SI	DBI	CHI	SI	DBI	CHI
Global	0.519	0.676	449.046	0.541	0.645	555.863
0	0.491	0.969	347.011	0.496	0.991	331.717
1	0.473	0.846	370.732	0.472	0.871	336.029
2	0.706	0.566	195.837	0.708	0.558	283.454

Pada Tabel 5 simulasi 2 terdapat *data train* (50%) dan *data test* (50%) untuk mengetahui simulasi terbaik maka dilakukan validasi menggunakan indeks SI, DBI, dan CHI. Validasi tersebut dilakukan pada keseluruhan *data train* dan *data test* dengan hasil perhitungan diberi label Global. Kemudian, validasi juga dilakukan pada masing-masing *cluster* yang mana hasil perhitungannya diberi label 0, 1, 2 sesuai dengan *cluster.* Untuk mempermudah membandingkan hasil dari setiap simulasi, maka dipilihlah hasil validasi dengan label Global, dimana nilai validasi dari masing-masing indeks pada *data train* dan *data test* dihitung nilai rata-ratanya sehingga menghasilkan nilai sebagai berikut0.530, 0.661, 502.454.

Untuk menentukan simulasi mana yang memiliki *performance* optimal dapat dilihat dari hasil evaluasi tersebut, dimana *score* dari silhouette index nya semakin mendekati 1, *score* dari davies-bouldin index nya kecil dan memiliki *score* calinski harabasz index yang besar [12].

4.3 Evaluasi

Pada tahap evaluasi akan dijelaskan mengenai hasil dari pengkajian dimana terdapat 3 model simulasi yaitu simulasi 1 (80%, 20%), simulasi 2 (70%, 30%), simulasi 3 (50%, 50%). Untuk menentukan simulasi yang memiliki *performance* yang optimal dilihat dari tiga indeks validitas dengan kriteria *relative*, yaitu indeks Silhouette, indeks Davies-Bouldin, dan Indeks Calinksi-Harabasz. Pada indeks Silhouette (SI) simulasi yang terbaik ditunjukkan dengan nilai silhouette yang semakin mendekati 1, Indeks validitas Davies-Bouldin simulasi terbaik ditunjukkan dengan nilai DBI yang semakin kecil, sedangkan untuk Indeks validitas Calinski-Harabasz (CHI) simulasi terbaik ditunjukkan dengan semakin besar nilai CHI [12].

Berdasarkan evaluasi *performance* yang telah dilakukan, dapat disimpulkan bahwa simulasi yang memiliki *performance* optimal merupakan simulasi 3 dikarenakan nilai validitas dari simulasi tersebut telah memenuhi kriteria validitas relative, dimana terdapat 2 kriteria yang sesuai yaitu memiliki nilai silhouette indeks mendekati 1 dan nilai CHI indeks dengan nilai yang semakin besar.

Hasil *clustering* dengan *data train* menunjukan *cluster* 0 merupakan pelanggan yang memiliki waktu berlangganan cukup lama dari mulai 7 – 11 bulan dengan jumlah layanan yang diambil mulai dari 1 – 4 layanan dengan total tagihan yang cukup variatif berkisar antara 286 ribu – 209 ribu**.** Sehingga pelanggan pada *cluster* 0 dikategorikan sebagai pelanggan yang memberikan keuntungan terbesar bagi perusahaan. *Cluster* 1 merupakan pelanggan yang memiliki waktu berlangganan 7, 8, dan 11 bulan dengan 1 jumlah layanan yang diambil dan dengan total tagihan paling rendah yaitu 275 ribu sedangkan total tagihan paling tinggi yaitu 1.144 ribu. Sehingga pelanggan pada cluster 1 dikategorikan sebagai pelanggan yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan. *Cluster* 2 merupakan pelanggan yang memiliki waktu berlangganan 1, 3, 4, dan 6 bulan dengan jumlah layanan yang diambil sebanyak 1 – 3 layanan dan dengan total tagihan paling rendah yaitu 286 ribu dan total tagihan yang paling tinggi yaitu 907.5 ribu. Sehingga pelanggan pada *cluster* 2 dikategorikan sebagai

pelanggan yang memberikan keuntungan lebih sedikit daripada biaya untuk memberikan pelayanan. Sedangkan *clustering* menggunakan *data test* memiliki hasil dimana *cluster* 0 merupakan pelanggan yang memiliki waktu berlangganan cukup lama dari mulai 7, 8, 10, dan 11 bulan dengan jumlah layanan yang diambil mulai dari 2 – 5 layanan dengan total tagihan yang cukup variatif berkisar antara 275 ribu – 209 ribu. Sehingga pelanggan pada *cluster* 0 dikategorikan sebagai pelanggan yang memberikan keuntungan terbesar bagi perusahaan. *Cluster* 1 merupakan pelanggan yang memiliki waktu berlangganan 7, 8, 10 dan 11 bulan dengan 1 jumlah layanan yang diambil dan dengan total tagihan paling rendah yaitu 275 ribu sedangkan total tagihan paling tinggi yaitu 1.094.5 ribu. Sehingga pelanggan pada *cluster* 1 dikategorikan sebagai pelanggan yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan. *Cluster* 2 merupakan pelanggan yang memiliki waktu berlangganan 1, 2, 3, 4, dan 6 bulan dengan jumlah layanan yang diambil sebanyak 1 – 4 layanan dan dengan total tagihan paling rendah yaitu 286 ribu dan total tagihan yang paling tinggi yaitu 1.518 ribu. Sehingga pelanggan pada *cluster* 2 dikategorikan sebagai pelanggan yang memberikan keuntungan lebih sedikit daripada biaya untuk memberikan pelayanan.

5. Kesimpulan

Kesimpulan yang dapat diambil dari penelitian Segmentasi Pelanggan Produk *Digital Service Add On* Indihome Menggunakan Algoritma K-Means yaitu hasil dari segmentasi data pelanggan menggunakan algoritma K-Means mempunyai nilai K = 3 sesuai dengan dengan hasil penentuan jumlah *cluster* terbaik menggunakan metode Elbow, sehingga data pelanggan tersebut dikelompokkan kedalam 3 segmen. Pada *customer profiling* data pelanggan dilihat dengan menganalisis anggota dari masing-masing *cluster* untuk mencari karakteristik setiap *cluster*nya. Hasil dari penelitian yaitu simulasi ke-3 dengan presentasi data train 50% dan data test 50% dimana customer profiling *cluster* 0 dengan kategori pelanggan yang memberikan keuntungan terbesar bagi perusahaan, *cluster* 1 pelanggan dengan kategori yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan, *cluster* 2 dengan kategori pelanggan yang memberikan keuntungan lebih sedikit dari pada biaya untuk memberikan pelayanan. Saran untuk pengembangan penelitian pada masa yang akan datang diantaranya perlu adanya implementasi dalam bentuk aplikasi ataupun API sehingga dapat memudahkan dalam melakukan proses segmentasi data pelanggan menggunakan algoritma K-Means.

Ucapan Terima Kasih

Ucapan terima kasih ditujukan kepada PT.Telekomunikasi Indonesia Regional III Jawa Barat yang telah memberikan izin untuk melakukan penelitian serta membantu proses pengumpulan data yang digunakan dalam penelitian.

Daftar Pustaka

[1] A. Z. Adnan, "Penerapan Strategi Promosi Pada Pemasaran Produk CV. Syntax Corporation Indonesia," *J. Ilm. Indones.*, vol. 3, no. 7, pp. 14–24, 2018, [Online]. Available: http://jurnal.syntaxliterate.co.id/index.php/syntax-literate/article/view/415. [Accessed: 18-Oct-2019].

[2] F. Nursa, H. Hardisman, and R. Semiarty, "Analisis Segmentasi dan Penentuan Target Pasar Pelanggan Instalasi Rawat Jalan Rumah Sakit Universitas Andalas," *J. Kesehat. Andalas*, vol. 8, no. 3, pp. 650–660, 2019, [Online]. Available: http://jurnal.fk.unand.ac.id/index.php/jka/article/view/1054. [Accessed: 18-Oct-

2019].

[3] V. R. Hananto, A. D. Churniawan, and A. P. Wardhanie, "Perancangan Analytical CRM untuk Mendukung Segmentasi Pelanggan di Institusi Pendidikan," *J. Ilm. Teknol. Inf. Asia*, vol. 11, no. 1, p. 79, 2017, [Online]. Available: https://jurnal.stmikasia.ac.id/index.php/jitika/article/view/55. [Accessed: 18-Oct-2019].

[4] C. Prianto and N. S. Harani, "The data mining analysis to determine the priorities of families who receiving assistance," *J. Phys. Conf. Ser.*, vol. 1280, no. 2, 2019, [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1280/2/022027/pdf. [Accessed: 18-Oct-2019].

[5] D. Triyansyah and D. Fitrianah, "Analisis Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing," *J. Telekomun. dan Komput.*, vol. 8, no. 3, p. 163, 2018, [Online]. Available: http://publikasi.mercubuana.ac.id/index.php/Incomtech/article/view/4174. [Accessed: 18-Oct-2019].

[6] B. E. Adiana, I. Soesanti, and A. E. Permanasari, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering," *Jutei*, vol. 2, no. 2, pp. 23–32, 2018, [Online]. Available: https://jutei.ukdw.ac.id/index.php/jurnal/article/view/76. [Accessed: 25-Oct-2019].

[7] A. A. Caraka, H. Haryanto, D. P. Kusumaningrum, S. Astuti, F. I. Komputer, and U. D. Nuswantoro, "Logika Fuzzy Menggunakan Metode Tsukamoto," *Techno.COM*, vol. 14, no. 4, pp. 255–265, 2015, [Online]. Available: http://publikasi.dinus.ac.id/index.php/technoc/article/view/970. [Accessed: 25-Oct-2019].

[8] F. E. M. Agustin, A. Fitria, and H. A. S, "Implementasi Algoritma K-Means Untuk Menentukan Kelompok Pengayaan Materi Mata Pelajaran Ujian Nasional (Studi Kasus: Smp Negeri 101 Jakarta)," *J. Tek. Inform.*, vol. 8, no. 1, pp. 73–78, 2015, [Online]. Available: http://journal.uinjkt.ac.id/index.php/ti/article/view/1938. [Accessed: 25-Oct-2019].

[9] Asroni and R. Adrian, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang," *Ilm. Semesta Tek.*, vol. 18, no. 1, pp. 76–82, 2015, [Online]. Available: http://journal.umy.ac.id/index.php/st/article/view/708. [Accessed: 25-Oct-2019].

[10] A. F. Harismawan, A. P. Kharisma, and T. Afirianto, "Analisis Perbandingan Performa Web Service Menggunakan Bahasa Pemrograman Python , PHP , dan Perl pada Client Berbasis Android," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 2, no. January, pp. 237–245, 2018, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/781. [Accessed: 25-Oct-2019].

[11] A. T. Rahman, Wiranto, and A. Rini, "Coal Trade Data Clustering Using K-Means (Case Study Pt. Global Bangkit Utama)," *ITSMART J. Teknol. dan Inf.*, vol. 6, no. 1, pp. 24–31, 2017, [Online]. Available: https://jurnal.uns.ac.id/itsmart/article/download/11296/11108. [Accessed: 25-Oct-2019].

[12] A. F. Khairati, A. A. Adlina, G. F. Hertono, and B. D. Handari, "Kajian Indeks Validitas pada Algoritma K-Means Enhanced dan K-Means MMCA," *Pros. Semin. Nas. Mat.*, vol. 2, pp. 161–170, 2019, [Online]. Available:

https://journal.unnes.ac.id/sju/index.php/prisma/article/download/28906/12636. [Accessed: 25-Oct-2019].

[13] H. Dhika and F. Destiawati, "Application of Data Mining Algorithm to Recipient of Motorcycle Installment," *ComTech Comput. Math. Eng. Appl.*, vol. 6, no. 4, p. 569, 2015, [Online]. Available: https://journal.binus.ac.id/index.php/comtech/article/view/2192. [Accessed: 01-Dec-2019].

[14] A. R. Condrobimo, A. V. D. Sano, and H. Nindito, "The Application Of K-Means Algorithm For LQ45 Index on Indonesia Stock Exchange," *ComTech Comput. Math. Eng. Appl.*, vol. 7, no. 2, p. 151, 2016, [Online]. Available: https://journal.binus.ac.id/index.php/comtech/article/view/2256. [Accessed: 01-Dec-2019].

[15] J. C. Manggala, "Tugas Akhir Implementasi GoBGP Sebagai Aplikasi Control Plan Pada Docker Container," Universitas Muhammadiyah Malang, 2019.

[16] W. M. P. Dhuhita, "Clustering Menggunakan Metode K-Mean Untuk Menentukan Status Gizi Balita," *J. Inform. Darmajaya*, vol. 15, no. 2, pp. 160–174, 2015, [Online]. Available: http://garuda.ristekdikti.go.id/documents/detail/568725. [Accessed: 01-Dec-2019].