Optimasi K-Means Dengan Particle Swarm Optimization Dalam Penentuan Titik Awal Pusat Klaster Data Telekomunikasi

K-Means Optimization with Particle Swarm Optimization In Determining The Starting Point Of Cluster Centers of Telecommunication Data

Raden Gesit Prasasti Alam[1], Yan Everhard[2]
[1,2]Fakultas Teknologi Informasi, Magister Ilmu Komputer , Universitas Budi Luhur, Jakarta, Indonesia
E-mail: [1] rgesit@gmail.com, [2]yan.everhard@budiluhur.ac.id

Abstrak

Berkembangnya kebutuhan masyarakat terhadap layanan telekomunikasi telah mengakibatkan persaingan yang semakin sengit di antara perusahaan-perusahaan di industri tersebut. Oleh karena itu, perusahaan-perusahaan ini dituntut untuk mencari strategi promosi guna meningkatkan penjualan produk mereka. SiDompul adalah aplikasi yang dikembangkan XL Axiata untuk membantu RO (Retail Outlet) dalam melakukan penjualan paket data XL Axiata. Di tahun 2022 terjadinya penurunan penjualan paket data XL Axiata dilihat dari transaksi penjualan Retail Outlet melalui aplikasi SiDompul mengalami penurunan 37% dan targetnya tidak tercapai. Penelitian ini bertujuan untuk membentuk dan menguji pemodelan klasterisasi dari data transaksi penjualan dengan metode K-Means dan metode Particle Swarm Optimization (PSO). PSO untuk optimasi penentuan pusat klaster atau centroid. Pada penelitian ini, Algoritma K-Means dan Particle Swarm Optimization (PSO) terbukti dapat membentuk 2 klaster yang lebih baik dimana nilai quantization errornya dan nilai SSE lebih rendah yaitu quantization 2.920 dan SSE 17.255 sedangkan pada K-Means quantization 2.939 dan SSE 17.288.

Kata kunci: Klasterisasi, Strategi Promosi, paket data telekomunikasi, K-Means, PSO

Abstract

The growing public demand for telecommunications services has resulted in increasingly fierce competition among companies in the industry. Therefore, these companies are required to find promotional strategies to increase the sales of their products.. SiDompul is an application developed by XL Axiata to assist RO (Retail Outlet) in selling XL Axiata data packages. In 2022, there was a decline in XL Axiata data package sales, as seen from Retail Outlet sales transactions through the SiDompul application, which decreased by 37% and the target was not achieved. This study aims to establish and test clustering modeling of sales transaction data with the K-Means and Particle Swarm Optimization (PSO) algorithms. PSO for optimizing the determination of the cluster center or centroid. In this study, the K-Means and PSO algorithms were proven to be able to form 2 better clusters where the quantization error values and SSE values were lower, namely quantization 2,920 and SSE 17,255 while KMeans quantization was 2,939 and SSE 17,288.

Keywords: Clusterization, Promotion Strategy, telecommunication data package, K-Means, PSO

1. PENDAHULUAN

Dalam era industri yang semakin terdigitalisasi, strategi promosi merupakan landasan penting bagi perusahaan untuk memperkenalkan produk dan layanan mereka kepada konsumen. XL Axiata, sebagai perusahaan telekomunikasi terkemuka di Indonesia, menghadapi tantangan

dalam mengembangkan strategi promosi yang efisien, terutama terkait penjualan paket data mereka. Dalam upaya mengatasi kendala ini, pendekatan klasterisasi menjadi fokus utama dalam memahami perilaku konsumen, preferensi produk, dan pola pembelian [1].

Melalui penerapan algoritma Klasterisasi K-Means serta pendekatan Particle Swarm Optimization (PSO), perusahaan berharap dapat mengoptimalkan penggunaan data transaksi penjualan untuk merumuskan strategi promosi yang lebih tepat sasaran dan berkelanjutan.

Analisis data transaksi penjualan menjadi kunci utama dalam mendapatkan pemahaman mendalam tentang perilaku konsumen. Algoritma Klasterisasi K-Means dipilih sebagai instrumen utama dalam mengelompokkan data transaksi penjualan paket data XL Axiata ke dalam klaster yang membagikan ciri-ciri tertentu. Dengan memisahkan data menjadi kelompok yang memiliki kesamaan, perusahaan dapat memperoleh wawasan yang lebih mendalam tentang preferensi konsumen, pola pembelian, dan segmentasi pasar yang beragam. Hal ini menjadi dasar untuk mengembangkan strategi promosi yang lebih terarah dan sesuai dengan kebutuhan masing-masing segmen pasar [2].

Dalam situasi ini, algoritma Klasterisasi K-Means dipilih sebagai instrumen utama untuk mengamati dan menggali wawasan berharga dari data transaksi penjualan. Algoritma K-Means ialah salah satu pendekatan klasterisasi partisional yang paling terkenal dan banyak dipakai karena kemudahan dalam analisis dan penerapannya, mampu menangani jumlah data yang besar, serta mempunyai proses yang relatif singkat. Namun, di samping keunggulannya, K-Means juga memiliki keterbatasan terkait penentuan titik awal pusat klaster yang dapat menyebabkan hasil klasterisasi kurang akurat. Selain itu, dalam proses pembaharuan titik pusat, terdapat potensi hasil klaster konvergen pada optima local [3].

Pada catatan berbeda, dalam upaya mengatasi kelemahan yang ditemukan dalam metode K-Means, yakni ketergantungan hasil klasterisasi pada penentuan pusat awal klaster, metode Particle Swarm Optimization (PSO) diterapkan untuk mengoptimalkan penentuan titik awal klaster. Pendekatan ini telah dijelaskan dalam penelitian terdahulu dan pemanfaatan PSO bersama K-Means telah membuktikan hasil yang lebih optimal dibandingkan dengan metode klasterisasi K-Means murni [4].

Penerapan teknik klasterisasi, khususnya dengan menggabungkan metode K-Means dan pendekatan Particle Swarm Optimization (PSO), diharapkan dapat memberikan kontribusi signifikan bagi XL Axiata. Dengan memanfaatkan data mining dan analisis klasterisasi, perusahaan dapat mengubah data transaksi penjualan menjadi pengetahuan yang berharga. Pemahaman mendalam tentang perilaku konsumen serta preferensi produk dari hasil klasterisasi ini diharapkan dapat mendukung perumusan strategi promosi yang lebih akurat, mengoptimalkan penargetan promosi, dan akhirnya, meningkatkan volume penjualan, bahkan melampaui target yang telah ditetapkan sebelumnya [5].

Dalam tinjauan studi dari penelitian sebelumnya terhadap metode klasterisasi dengan beberapa algoritma perbandingannya diantaranya dengan menggunakan beberapa metode klastering diantaranya K-Means, K-Medoids, Fuzzy C-Means, X-Means, DBSCAN dan FP-Growth. Dilihat dari hasil penelitian sebelumnya akurasi dengan algoritma K-Means lebih unggul dari algoritma lainnya diantaranya penelitian dari [6]yang melakukan riset untuk menggolongkan data pelanggan berdasarkan kisaran waktu dan jumlah pembelian menggunakan algoritma klasterisasi, yakni K-means, K-medoids, dan Fuzzy C-means.

Hasil studinya menunjukkan bahwa algoritma K-Means lebih efisien daripada K-Medoids dan Fuzzy C-means dalam pengklasteran data pelanggan, yang dibuktikan melalui skor validitas DBI terbaik mencapai 0,167 dengan jumlah klaster 6. Selain itu, dalam perbandingan K-Means dengan DBSCAN, K-Means lebih superior dalam mengklasterisasi kasus Covid-19. Algoritma K-Means mendapatkan nilai SI optimal sebesar 0,6902 [7].

Kemudian untuk membantu mengoptimalisasi dalam menentukan pusat klaster di K-Means kita akan mengkombinasikan dengan metode Particle Swarm Optimization (PSO) yang lebih stabil nilai fitness dibandingkan dengan metode GA (Genetic Algorithm) yaitu sebesar 0.111 [8] termasuk menggunakan metode pengujian CHI (Calinski-Harabasz Index) salah satunya yang efektif digunakan untuk menentukan kualitas dalam sebuah segmentasi kelompok [9].

Dari Beberapa hasil penelitian dan perbandingan sebelumnya, penelitian ini menentukan dengan algoritma Klasterisasi K-Means untuk menyempurnakan penelitian sebelumnya terhadap klasterisasi data penjualan paket data dan akan divalidasi dengan DBI (Davies Bouldin Index) dan Silhouette. Termasuk dilakukan improvisasi dan perbaikan akurasi terhadap penelitian sebelumnya yang hanya menentukan data penjualan paket data [10] namun dengan penelitian ini mengoptimalisasi penentuan titik pusat klaster K-Means dengan Metode Particle Swarm Optimization (PSO) dan melakukan pengujian hasil klasterisasi dengan 4 metode pengujian secara komprehensif yaitu Elbow SSE, CHI (Calinski-Harabasz Index), DBI (Davies Boldin Index) dan Silhouette untuk membuktikan jumlah klaster yang terbaik dari semua pengujian sekaligus menambahkan banyak atribut dan variabel datanya untuk membentuk klaster terbaik dalam menunjang promosi penjualan paket data.

2. METODE PENELITIAN

2.1. Metodologi Penelitian

Metodologi penelitian dalam *data science* mencakup langkah-langkah sistematis yang dilakukan untuk merencanakan, melaksanakan, dan mengevaluasi studi yang berkaitan dengan data. Dalam penelitian ini, *model data mining* yang digunakan adalah *Cross Industry Standard Process for Data mining* atau yang disingkat CRISP-DM.

CRISP-DM (*Cross-Industry Standard Process for Data Mining*) adalah metodologi yang umum digunakan dalam proses pengembangan dan analisis data. Metodologi ini memberikan panduan langkah demi langkah untuk mengelola proyek *data science* dari awal hingga akhir.

Berikut adalah konsep utama dalam CRISP-DM:

a) Pemahaman Bisnis (*Business Understanding*)

Langkah pertama dalam CRISP-DM adalah memahami masalah bisnis dan tujuan yang ingin dicapai melalui analisis data. Hal ini melibatkan identifikasi tujuan bisnis, kebutuhan pengguna, serta memahami konteks dan batasan proyek.

b) Pemahaman Data (*Data Understanding*)

Langkah selanjutnya adalah memahami data yang tersedia untuk analisis. Ini termasuk mengumpulkan data, mengeksplorasi dan mengidentifikasi karakteristik data, serta mengevaluasi kualitas data. Tujuannya adalah memahami data dengan lebih baik sebelum melakukan analisis.

c) Persiapan Data (*Data Preparation*)

Pada langkah ini, yang pertama data dipersiapkan untuk analisis lebih lanjut. Ini melibatkan pembersihan data, transformasi, integrasi data dari berbagai sumber, serta pemilihan variabel yang relevan. Tujuan dari tahap ini adalah memastikan data siap digunakan untuk membangun model dan melakukan analisis.

d) *Modelling*

Pada langkah ini, model atau teknik analisis data dikembangkan untuk menjawab pertanyaan bisnis atau mencapai tujuan proyek. Ini melibatkan pemilihan teknik/model yang tepat, pembangunan model, serta pengujian dan validasi model. Pada akhir tahap ini, model yang dihasilkan harus dapat memberikan wawasan yang berarti.

e) Evaluasi

Setelah membangun model, langkah selanjutnya adalah mengevaluasi performa dan kualitas model. Ini melibatkan analisis hasil model, pengukuran performa, serta mengevaluasi apakah model memenuhi tujuan bisnis dan kriteria kesuksesan proyek.

f) Penyampaian (*Deployment*)

Tahap ini melibatkan penerapan hasil analisis dalam lingkungan produksi atau penggunaan praktis. Model yang dihasilkan diterapkan untuk mengambil keputusan bisnis, atau hasil analisis disajikan dalam bentuk yang dapat dimengerti oleh pengguna. Proses ini juga melibatkan dokumentasi dan komunikasi hasil kepada stakeholder terkait.

g) Siklus Iterasi

CRISP-DM adalah metodologi siklus hidup yang iteratif. Oleh karena itu, setelah tahap penyampaian, hasil evaluasi dan penggunaan dapat digunakan untuk memperbaiki atau memperluas model yang ada, atau untuk mengarahkan analisis data berikutnya.
CRISP-DM memberikan pendekatan sistematis dalam mengelola proyek *data science*, dengan membagi proses menjadi langkah-langkah yang jelas dan terstruktur. Metodologi ini membantu mengurangi risiko dan memastikan bahwa analisis data yang dilakukan relevan dengan tujuan bisnis yang ingin dicapai.

2.1.1. *Business Understanding*

Pada business understanding ini mengacu kepada permasalahan yang ada. Permasalahan yang ada tersebut adanya penurunan transaksi penjualan paket data XL dan tidak mencapai target. Penurunan yang terjadi sekitar 37% pada tahun 2022. Permasalahan ini berada pada bagian *Sales* dan distribusi. Maka dari itu, pada tahapan ini diperlukan pemahaman tentang pentingnya pemanfaatan dan pengolahan data transaksi penjualan agar dapat digunakan untuk pengelompokan segmen pelanggan dari masing-masing *region*, kota dan atribut lainnya dengan tujuan setiap klaster yang terbentuk dengan optimasi metode PSO menjadi lebih akurat dan dapat menjadikan dasar untuk bagian *Sales* dan distribusi untuk melakukan strategi promosi. Sekaligus untuk mengefisiensikan bentuk promosi yang dilakukan tim *sales* dan distribusi baik itu promosi *digital* ataupun *non digital*/tradisional bisa lebih akurat menyasar pelanggan/*segment* yang lebih spesifik untuk bisa menaikkan *target* penjualan paket data XL.

2.1.2. *Data Understanding*

Data Understanding adalah proses yang mempertemukan antara data yang dimiliki dan data yang seharusnya perlukan. Dataset yang digunakan pada penelitian ini dikumpulkan dari bagian penjualan dan distribusi PT. XL Axiata. Data yang digunakan adalah data sampel penjualan selama 3 tahun 2020-2022 sebanyak 2.495 data. Tidak semua atribut data dari transaksi penjualan digunakan, hanya memilih atribut yang diantaranya Nama RO (Retail Outlet), Lokasi Region, Kota pembelian, Produk/Paket data yang dibeli, Jumlah transaksi dan Harga paket data.

2.1.3. *Data Preparation*

Proses persiapan data adalah langkah penting dalam memperlakukan data menuju model yang berkualitas dan bermanfaat. Pada tahap ini, dilakukan pengolahan awal data, pemilihan atribut atau variabel yang akan dianalisis, validasi variabel yang ada, dan persiapan transformasi data. Hasil yang dihasilkan dari tahap persiapan data ini mencakup :

1. Melakukan Data *Cleaning* terhadap data yang memiliki *missing value* (*isNull*). Jika masih terdapat data yang *missing value* maka akan dikeluarkan dari dataset.
2. Data *Grouping* berdasarkan *Region* dan nama paket data untuk mengelompokkan data secara spesifik melihat sebaran data berdasarkan masing-masing *region* dan paket data yang terjual
3. Dilakukan seleksi variabel yang akan dianalisis dan validasi variabel yang digunakan dalam penelitian ini terteta pada tabel 1
4. Melakukan Normalisasi data untuk menjaga jarak masing-masing data tidak terlalu jauh sehingga merusak sebaran klaster dan mempunyai skala data yang standar.

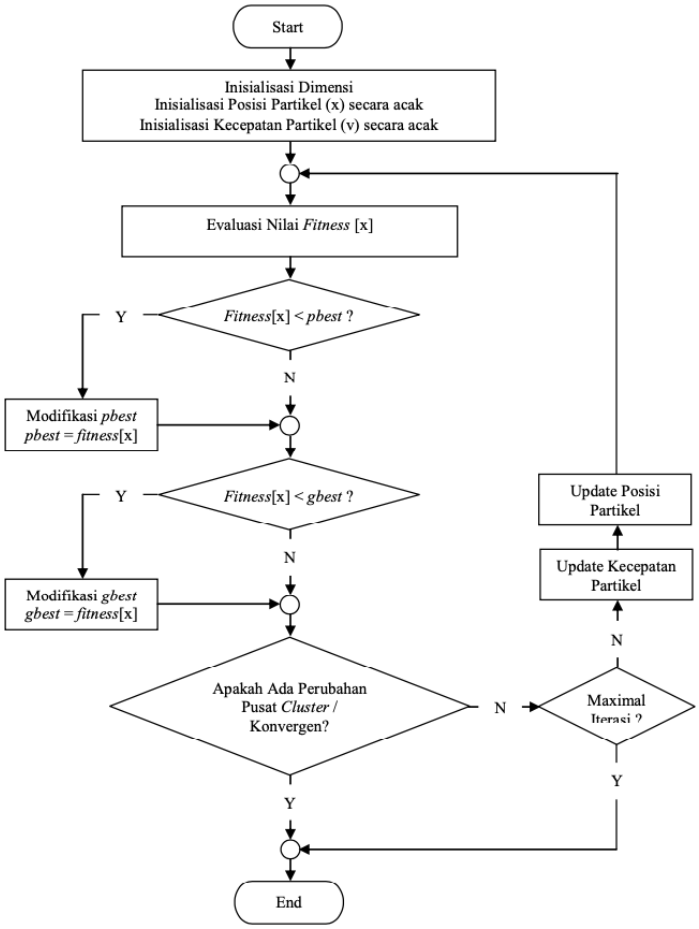
Tabel 1 Data atribut transaksi penjualan

No	Nama Atribut	Keterangan	Alasan
1	Nama RO (*Retail Oulet*)	- Toko A - Toko B	Dilakukan perhitungan frekuensi Nama-nama RO yang sering transaksi penjualan
2	Region	- Pusat - Timur - Barat	Pembagian Lokasi dari RO akan dilakukan perhitungan frekuensi Regionnya

3	Kota	- Semarang - Cianjur - Pamekasan - Surabaya	Frekuensi perhitungan kota yang sering terjadi transaksi penjualan
4	Paket Data	- Paket 3GB - Paket 5GB - Paket 24 Jam 10GB	Untuk mnegetahui paket/produk mana yang sering dibeli
5	Jumlah Transaksi	-1 -5 -10	Mengetahui jumlah transaksi dari setiap pembelian produk
6	Jumlah Pendapatan	Nilai mata uang (Rupiah)	Pendapatan untuk XL Axiata dari setiap penjualan melalui *Retail Outlet*
7	Harga paket data	Nilai mata uang (Rupiah)	Harga satuan setiap paket data yang ditawarkan kepada pelanggan

2.1.4. *Modeling*

Dalam fase pemodelan data menggunakan algoritma klasterisasi K-Means sekaligus mengoptimalisasi penentuan titik pusat klaster K-Means dengan Metode *Particle Swarm Optimization* (PSO) dan melakukan pengujian hasil klasterisasi dengan 4 metode pengujian secara komprehensif. Tahapan PSO ditunjukkan pada gambar 1.



Gambar 1 Proses Metode Particle Swarm Optimization (PSO) untuk penentuan Pusat Klaster

2.1.5. *Evaluation*

Tahap kelima adalah Penilaian. Setelah model atau beberapa model telah diperoleh, dilakukan evaluasi terhadap kualitas dan efektivitasnya.

Evaluasi dengan menggunakan beberapa metrik untuk pengujian klasterisasi. Dengan menggunakan beberapa metrik untuk mengukur kualitas klasterisasi, adapun metrik yang digunakan adalah *Elbow, Silhouette Score, Calinski-Harabasz Index dan Davies-Bouldin Index* (DBI)

2.1.6. *Deployment*

Tahapan akhir dalam model CRISP-DM adalah Penerapan. Persiapan untuk Penerapan dimulai sejak Pemahaman Bisnis dan harus mempertimbangkan tidak hanya cara menghasilkan nilai dari model, melainkan juga konversi skor keputusan serta integrasi keputusan ke dalam sistem operasional.

Program *prototype* ini diimplementasikan menggunakan pendekatan bahasa pemrograman *Python* dan memanfaatkan diantaranya library seperti Flask, Pandas, dan pickel untuk menyimpan dan membaca pemodelan dari KMeans klasterisasi. Flask digunakan sebagai *framework web* untuk membangun aplikasi web sederhana, Pandas digunakan untuk manipulasi dan analisis data. Di akhirnya, rencana penyebaran sistem mengakui bahwa tidak ada model yang tetap. Model ini dibentuk dari data yang mencerminkan kondisi pada waktu tertentu, sehingga perubahan waktu bisa mengubah karakteristik data. Oleh karena itu, model juga perlu diawasi dan digantikan oleh model yang telah diperbaiki.

2.2. Sampling/Metode Pemilihan Sampel

Dalam *data science*, pemilihan sampel data yang tepat sangat penting untuk menghasilkan hasil analisis yang akurat dan dapat dipercaya. Sampel secara acak Sederhana (*Simple Random Sampling*): Metode ini mirip dengan yang digunakan dalam statistik. Anda dapat menggunakan fungsi acak atau algoritma acak dalam bahasa pemrograman untuk memilih sampel secara acak dari data Anda. Pastikan setiap data memiliki peluang yang sama untuk dipilih.

Sampel Proporsional (*Proportional Sampling*): Metode ini digunakan ketika Anda ingin mempertahankan proporsi tertentu dalam sampel yang mewakili proporsi di populasi asli. Misalnya, jika Anda memiliki kelas atau kelompok dalam data Anda, Anda dapat memilih proporsi yang sama dari setiap kelas atau kelompok untuk memastikan representativitas sampel.

Sampel Stratifikasi (*Stratified Sampling*): Metode ini melibatkan pemilihan sampel dari setiap stratum atau kelompok yang ada dalam data. Dalam stratifikasi, Anda membagi data ke dalam kelompok-kelompok berdasarkan karakteristik tertentu, dan kemudian memilih sampel dari setiap kelompok. Hal ini memungkinkan Anda memastikan bahwa setiap kelompok atau kelas dalam data Anda terwakili dengan baik dalam sampel.

Sampel Berbasis Cluster (*Cluster Sampling*): Metode ini melibatkan pemilihan sampel dengan memilih secara acak beberapa kluster atau kelompok dari data Anda dan menggunakan semua data dalam kluster yang dipilih sebagai sampel. Metode ini efisien ketika data Anda terstruktur dalam kluster atau kelompok yang dapat diperoleh dengan mudah.

Sampel Berbasis Waktu (*Time-based Sampling*): Metode ini cocok ketika Anda memiliki data berurutan berdasarkan waktu, seperti data yang dihasilkan dalam rentang waktu tertentu. Anda dapat memilih sampel dengan mengambil titik data dalam interval waktu yang ditentukan, atau dengan memilih titik data pada interval waktu yang teratur.

Sampel Berbasis Probabilitas (*Probability-based Sampling*): Metode ini melibatkan penggunaan probabilitas dan model statistik untuk memilih sampel. Ini sering digunakan dalam teknik pengambilan sampel yang lebih kompleks, seperti sampel berstrata berlapis atau sampel dengan probabilitas terpengaruh oleh karakteristik tertentu.

Pemilihan metode sampling yang tepat dalam *data science* tergantung pada jenis data yang Anda miliki, tujuan analisis, dan kendala yang ada. Selain itu, penting untuk memastikan

bahwa sampel yang Anda pilih mencerminkan populasi asli dan mewakili variasi yang ada dalam data.

Dataset yang digunakan merupakan data transaksi penjualan paket data XL melalui aplikasi SiDompul pada periode 3 tahun yaitu dari tahun 2020 - 2022. Dataset transaksi penjualan terdiri kolom/atribut yang terdiri dari Nama RO (*Retail Outlet*), Lokasi *Region*, Kota pembelian, Produk/Paket data yang dibeli, Jumlah transaksi dan Harga paket data. Pengambilan data sampel ini menggunakan sampel berbasis region sehingga semua data dapat terwakili selama 12 bulan setiap tahunnya. Dataset sampel yang berjumlah 2.495 tersebut dilakukan ujicoba dengan berbagai data training dan data testing dengan algoritma KMeans dan dioptimasi dengan metode *Particle Swarm Optimization* (PSO) dan diuji data dengan 4 metode pengujian.

2.3. Metode Pengumpulan Data

Metode perolehan data merupakan langkah yang sangat penting dalam proses penelitian karena tujuan utama dari penelitian adalah memperoleh data. Untuk mencapai tujuan ini, digunakan berbagai teknik pengumpulan data, yaitu:

1. Metode Observasi

Observasi adalah teknik perolehan data yang melibatkan pengamatan dan pencatatan yang teliti dan sistematis terhadap situasi atau fenomena yang sedang diselidiki. Peneliti melakukan pengamatan langsung di Perusahaan XL Axiata untuk mengidentifikasi permasalahan serta memahami sistem dan proses yang diterapkan di perusahaan tersebut.

2. Metode Wawancara

Wawancara adalah teknik pengumpulan data yang dilakukan melalui interaksi tatap muka dan tanya jawab langsung antara peneliti dan narasumber. Dengan perkembangan teknologi, wawancara juga bisa dilakukan melalui panggilan daring, email, dan pertemuan menggunakan *platform* seperti *Microsoft Teams*. Peneliti melakukan tanya jawab langsung dengan *Product Owner* dan *Growth Hacker* XL Axiata. Hasil wawancara ini memberikan tambahan informasi yang melengkapi hasil observasi.

3. Metode Dokumentasi

Dokumentasi adalah cara untuk memperoleh dokumen-dokumen yang mencakup bukti akurat dari catatan informasi tertulis, seperti buku atau sumber lainnya. Dalam penelitian ini, pengumpulan data melalui dokumen dilakukan dengan menganalisis fakta atau informasi yang terdapat dalam berkas dokumen transaksi penjualan di PT. XL Axiata.

2.4. Instrumentasi

Instrumentasi dalam penelitian ini diantaranya menggunakan perangkat keras (*hardware*), perangkat lunak (*Software*) dan instrument untuk studi dokumentasi

1. Perangkat Keras (*Hardware*)
    a. MacBook Pro 13 Inch 2017
    b. Processor 2,3 GHz Dual-Core Intel Core i5
    c. Memory 8 GB 2133 MHz LPDDR3
2. Perangkat Lunak (*Software*)
    a. MacOS Ventura 13.4.1
    b. Bahasa pemrograman Python Versi 3.10.8
    c. Visual Studio Code with Jupiter notebook

2.5. Teknik Analisis, Perancangan dan Pengujian Data/*Prototype Model*

2.5.1. Teknik Analisis

Teknik analisis yang dilakukan adalah dengan melakukan klastersasi terhadap dataset 3 tahun dari tahun 2020 - 2022 dengan tambahan atribut dan *variable* dengan algoritma K-Means *Clustering* yang dioptimasi dengan metode *Particle Swarm Optimization* (PSO). Kemudian melihat percobaan dari beberapa pemodelan klasterisasi yang terbentuk. Dari hasil pemodelan klasterisasi yang terbentuk tersebut akan diuji untuk mendapatkan akurasi terbaik.

2.5.2. Teknik Pengujian

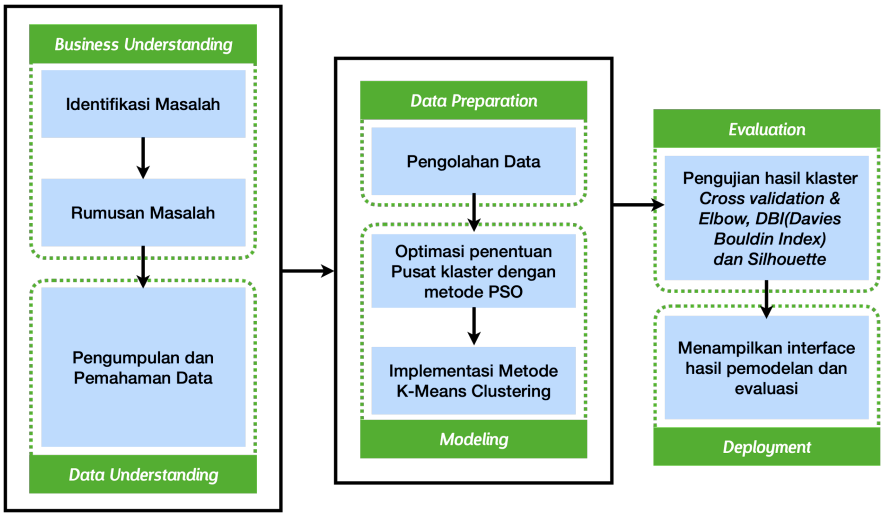
Pengujian terhadap hasil pemodelan klasterisasi yang diterapkan dalam penelitian ini adalah dengan 4 metode secara komprehensif yaitu *Elbow SSE*, *Calinski-Harabasz Index*, DBI (*Davies Boldin Index*) dan *Silhouette*.
Dari beberapa pengujian itu akan dilihat bagaimana akurasi tertinggi dari setiap klustering yang dibentuk.

2.5.3. *Prototype Model*

Prototype yang digunakan untuk melakukan deployment data model yang sudah diuji dan mendapatkan akurasi skor untuk di implementasikan. Proses *prototype* dimulai dengan menyimpan hasil klasterisasi model terbaik yang sudah diolah dengan menggunakan bahasa pemrograman *Python* kedalam bentuk grafik pengujian, dari data itu bisa ditampilkan atau ditambahkan menjadi API ke dalam aplikasi penjualan untuk ditampilkan dan dipelajari *user*/*distributor*.

2.6. Langkah-langkah Penelitian

Guna menerapkan sebuah penelitian, diperlukan langkah-langkah yang harus dijalankan agar akhirnya penelitian dapat direalisasikan dan menghasilkan hasil yang tepat. Proses penelitian yang dijalankan oleh peneliti bisa diamati pada Gambar 2.



Gambar 2 Langkah – langkah Penelitian

Dari Gambar 2 tersebut dapat dijelaskan tahap-tahap penelitian sebagai berikut :

1. Tahap Identifikasi Masalah
   Pada tahap Identifikasi masalah ini, peneliti melakukan identifikasi masalah dalam mencari kluster terbaik dari data transaksi untuk meningkatkan penjualan untuk strategi promosi yang dilakukan oleh bagian Penjualan dan Distribusi.
2. Tahap Rumusan Masalah
   Untuk merumuskan permasalahan yang terjadi di bagian Penjualan dan Distribusi PT. XL Axiata untuk mencari solusi yang akan digunakan.
3. Tahap Pengumpulan dan Pemahaman Data
   Yaitu mengumpulkan data dan melakukan beberapa metide diantaranya metode observasi, metode wawancara dan metode dokumentasi untuk mendapatkan data sebagai *literature review* melalui jurnal-jurnal yang sama untuk menganalisa dan menguji data sudah sesuai.

4. Tahap Pengolahan Data
   Pada tahap ini, peneliti mencari klaster terbaik penjualan terbaik untuk digunakan sebagai strategi promosi bagian penjualan dan distribusi
5. Tahap Implementasi metode
   Tahap ini menggunakan algoritma K-Means *Clustering* yang di-optimasi menggunakan metode *Particle Swarm Optimization* (PSO) dan implementasi dengan Bahasa pemrograman Python menggunakan library yang dibutuhkan untuk *Machine Learning* untuk mengimplementasikan data yang sudah diolah di tahap sebelumnya
6. Tahap Pengujian Data
   Tahap ini melakukan pengujian dengan menggunakan *Elbow*, *Calinski-Harabasz Index*, DBI (*Davies Boldin Index*) dan *Silhouette*.
7. *Deployment*
   Tahapan ini akan dilakukan penyajian *interface* dengan pemodelan yang sudah dibentuk dan di-*compile* sebelumnya.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, tujuan utama adalah untuk menerapkan algoritma k-means dengan optimasi penentuan pusat klaster dengan metode Particle Swarm Optimization (PSO) dalam melakukan klasterisasi data strategi promosi dari transaksi penjualan. Penerapan ini bertujuan untuk membuat kelompok data promosi ke dalam bentuk kelompok yang serupa berdasarkan pola dan karakteristik yang ada.
Berikut adalah tahapan penerapan model klaster dengan K-Means menggunakan *Python*:

a. Persiapan Data
Mempersiapkan data transaksi penjualan yang akan digunakan untuk klasterisasi strategi promosi. Melakukan pemrosesan data seperti penghapusan data yang tidak relevan, penanganan *missing value*, atau transformasi data jika diperlukan.

b. *Preprocessing* Data
Jika diperlukan, lakukan *preprocessing* data seperti normalisasi atau standardisasi data untuk memastikan bahwa setiap fitur atau atribut memiliki skala yang serupa.

c. Menentukan Jumlah Klaster
Tentukan jumlah klaster yang akan digunakan dalam klasterisasi. Dalam hal ini bisa dengan melakukan proses perhitungan menggunakan metode yang tepat seperti *Elbow Method, Silhouette Score, Davies Bouldin, Calinski Index*, atau melalui pemahaman domain yang baik.

d. Inisialisasi Model
Inisialisasi model K-Means dengan jumlah klaster yang telah ditentukan. Dengan penentuan Pusat klaster/*centroid* mengkombinasi dengan *Particle Swarm Optimization* (PSO)

e. Pelatihan Model
Latih model K-Means dengan menggunakan data transaksi penjualan yang telah diproses dan di-*preprocessed*.

f. Klasterisasi Data
Gunakan model K-Means yang telah dilatih untuk melakukan klasterisasi pada data strategi promosi.
Setiap data promosi akan diberikan label klaster berdasarkan kelompok yang paling sesuai.

g. Evaluasi Klasterisasi
Evaluasi hasil klasterisasi menggunakan metrik evaluasi seperti *Elbow Method, Silhouette Score, Davies Bouldin, Calinski Index* dalam penelitian ini.
Analisis hasil klasterisasi dan perhatikan apakah hasil klasterisasi sesuai dengan harapan dan tujuan penelitian.

h. Interpretasi Hasil Klasterisasi
Interpretasikan hasil klasterisasi untuk mendapatkan wawasan tentang kelompok strategi promosi yang serupa.
Analisis karakteristik atau pola yang ada di setiap klaster untuk pemahaman yang lebih baik tentang strategi promosi yang efektif.

Penerapan model klaster dengan K-Means dan Particle Swarm Optimization (PSO) menggunakan *Python* akan memberikan pemahaman yang lebih mendalam tentang strategi promosi berdasarkan analisis data penjualan. Proses ini dapat membantu dalam pengembangan strategi promosi yang lebih terfokus dan efektif untuk meningkatkan penjualan paket data.

Penelitian ini untuk mencari hasil jumlah klaster terbaik, dimana hasil dari uji coba pemodelan dari metode K-Means dan Particle Swarm Optimization (PSO) dengan menggunakan *python scripting* yang diukur berdasarkan jarak minimun terhadap *centroid* dengan maksimum iterasi = 100 dan jarak yang paling dekat atau terendah yang akan diambil dalam implementasi sistem.

Evaluasi Hasil Model Klastering dengan K-Means dan Particle Swarm Optimization (PSO)

Setelah melakukan penerapan algoritma K-Means dan Particle Swarm Optimization (PSO) dalam klasterisasi menggunakan contoh data, langkah selanjutnya adalah melakukan evaluasi hasil model dan pengujian. Evaluasi ini bertujuan untuk menilai kualitas klasterisasi yang telah dilakukan dan melihat sejauh mana model K-Means dapat mengelompokkan data strategi promosi dengan baik.

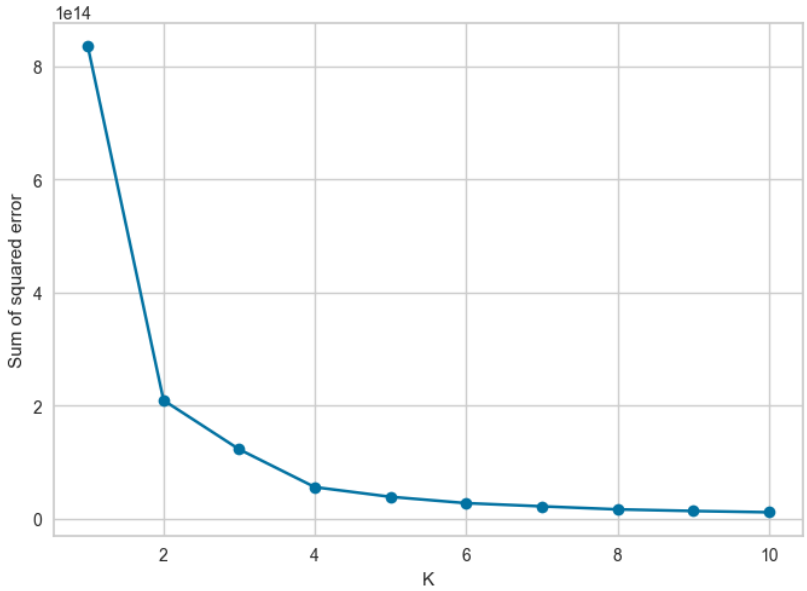
Terdapat beberapa langkah yang dilakukan untuk evaluasi hasil model dan pengujian:
Evaluasi dengan menggunakan beberapa metrik untuk pengujian klasterisasi dengan menggunakan beberapa metrik untuk mengukur kualitas klasterisasi, adapun metrik yang digunakan adalah *Elbow, Silhouette Score, Calinski-Harabasz Index* dan *Davies-Bouldin Index* (DBI)

Tabel 2 Hasil pengujian klasterisasi (dari Jumlah Klaster)

Uji coba (jumlah klaster)	Hasil pengujian dengan beberapa metrik			
	Elbow SSE (Sum of Squared error)	*Silhouette Score*	*Davies-Bouldin Index*	*Calinski-Harabasz Index*
2	**2.09**	**0.9341**	**0.3421**	**1011.30**
3	1.22	0.9261	0.2605	979.56
4	0.56	0.8567	0.3980	1561.21
5	0.38	0.8396	0.5278	1716.65
6	0.27	0.8398	0.5424	1947.91
7	0.22	0.8395	0.5280	2053.80
8	0.16	0.6606	0.5054	2324.70
9	0.13	0.6588	0.5867	2406.76
10	0.11	0.6600	0.5351	2679.04

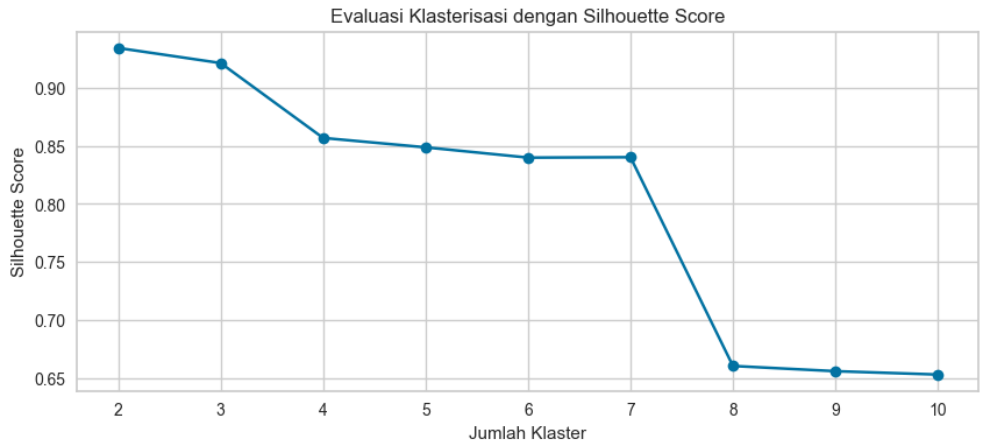
Dari Tabel 2, terdapat hasil pengujian klasterisasi menggunakan beberapa metrik untuk berbagai jumlah klaster yang berbeda. Berikut adalah keterangan dari hasil analisa tersebut:

1. *Elbow SSE (Sum of Squared Error*):
Elbow SSE merupakan metode yang digunakan untuk penentuan jumlah klaster optimal di dalam metode k-Means. Nilai *Elbow SSE* menurun ketika jumlah klaster meningkat. Dalam Gambar 3 terlihat bahwa nilai *Elbow SSE* menurun secara signifikan saat jumlah klaster bertambah dari 2 hingga 4, tetapi setelah itu penurunan menjadi lebih lambat. Pada titik 2 hingga 4, penurunan *Elbow SSE* mulai merata. Oleh karena itu, titik *elbow* atau siku terletak di sekitar jumlah klaster 2 hingga 4, menunjukkan bahwa 2 hingga 4 klaster merupakan pilihan yang baik.



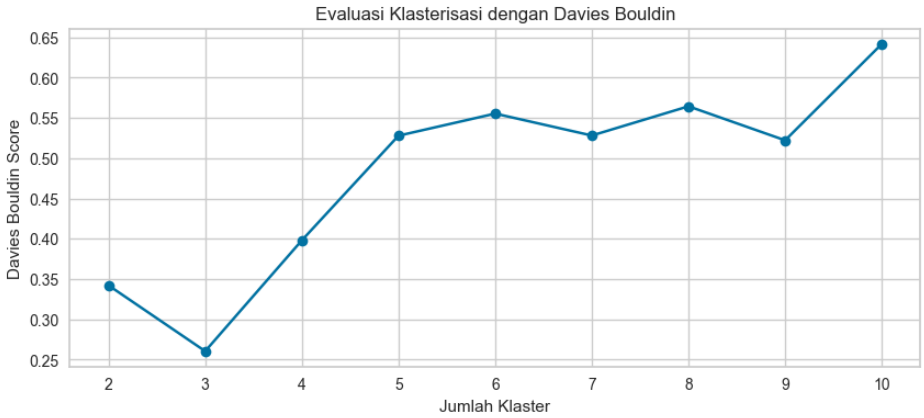
Gambar 3 Hasil pengujian Elbow - SSE

2. *Silhouette Score*:
Silhouette Score adalah metrik evaluasi yang mengukur sejauh mana setiap sampel cocok dengan klasternya dan sejauh mana klaster tersebut terpisah dari klaster lainnya. Semakin tinggi nilai *Silhouette Score*, semakin baik kualitas klasterisasi. Dalam Gambar 4 nilai *Silhouette Score* terlihat tinggi dan stabil dalam rentang 0.83 hingga 0.93, menunjukkan kualitas klasterisasi yang baik.



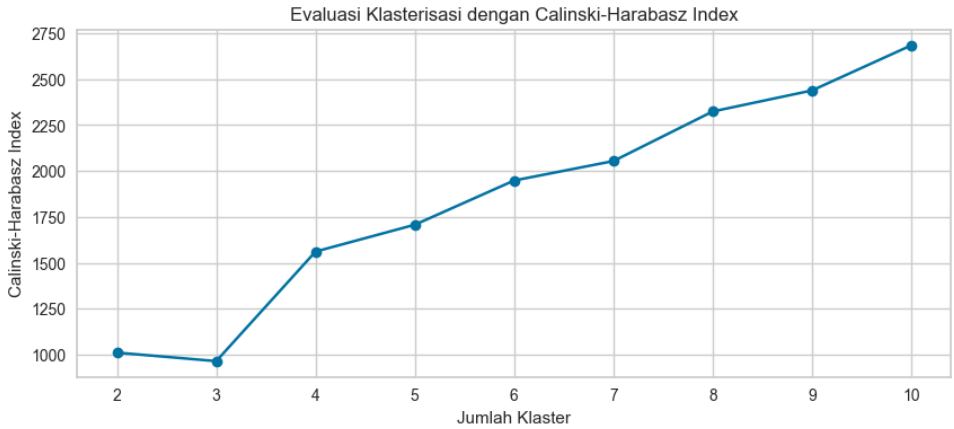
Gambar 4 Hasil pengujian dengan Silhouette Score

3. *Davies-Bouldin Index*:
Davies-Bouldin Index (DBI) adalah metrik evaluasi untuk melakukan pengukuran terhadap kualitas klasterisasi berdasarkan jarak antara klaster dan kepadatan klaster. Semakin rendah nilai DBI, semakin baik kualitas klasterisasi. Dalam Gambar 5 nilai DBI terlihat rendah dan stabil dalam rentang 0.26 hingga 0.59, menunjukkan bahwa klasterisasi memiliki kepadatan yang baik dan klaster yang terpisah dengan jarak yang cukup.



Gambar 5 Hasil pengujian dengan Davies Bouldin

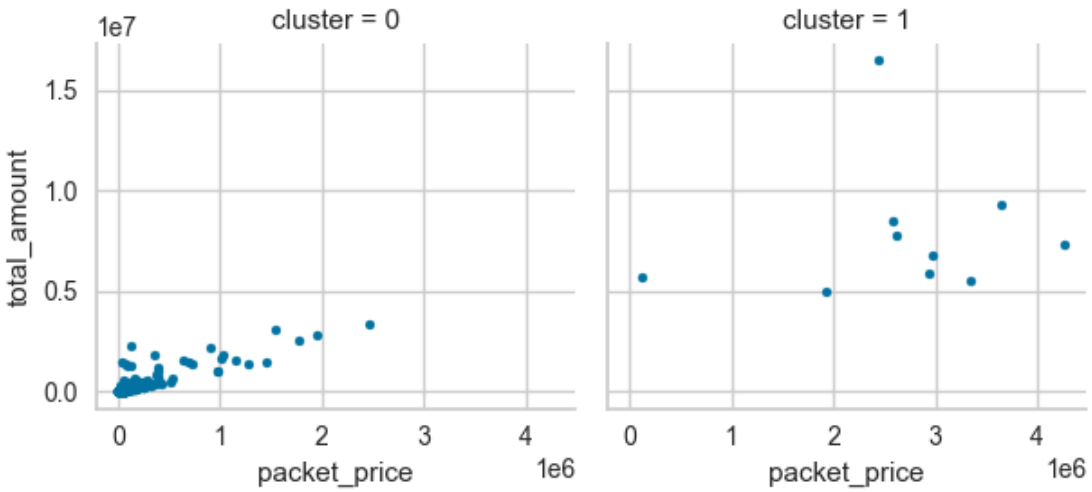
4. *Calinski-Harabasz Index*:
Calinski-Harabasz Index (CHI) adalah metrik evaluasi yang mengukur kualitas klasterisasi berdasarkan perbedaan antara varian dalam klaster dengan varian antar klaster. Semakin tinggi nilai CHI, semakin baik kualitas klasterisasi. Dalam Gambar 6 nilai CHI terlihat tinggi dan meningkat seiring dengan peningkatan jumlah klaster, menunjukkan bahwa klasterisasi memiliki perbedaan varian yang signifikan antara klaster. Namun dikarenakan dari jarak klaster ke 2 dan ke 3, dimana klaster ke 3 mengalami penurunan. Jadi pemilihan klaster ke 2 di awal bisa lebih optimal berdasarkan metrik CHI dilihat dari jarak klaster 2 ke 3



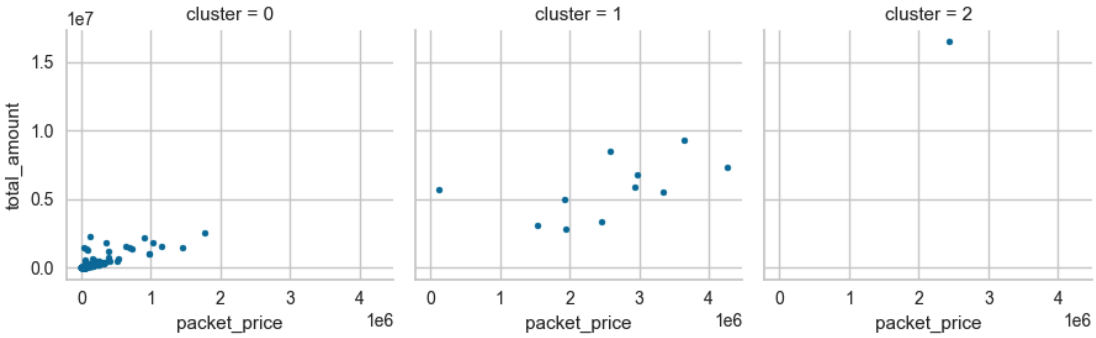
Gambar 6 Hasil pengujian dengan Calinski-Harabasz Index

Berdasarkan hasil analisis ini, dapat disimpulkan bahwa jumlah klaster 2 sampai 4 merupakan pilihan yang baik berdasarkan metrik *Elbow SSE* dan karakteristik yang diperlihatkan oleh metrik lainnya seperti *Silhouette Score, Davies-Bouldin Index,* dan *Calinski-Harabasz Index*. Terlampir hasil sebaran data dengan 2 klaster di Gambar 7, hasil sebaran data dengan 3 klaster di Gambar 8 dan hasil sebaran data dengan 4 klaster di Gambar 9. Namun, pemilihan jumlah klaster yang

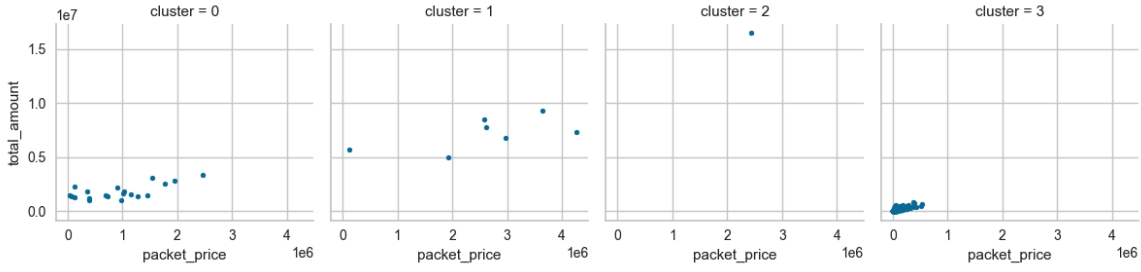
optimal juga perlu mempertimbangkan konteks dan tujuan analisis yang lebih spesifik. Dan karena klastering awal digunakan untuk strategi promosi untuk akselerasi cepat penulis memilih klaster ke 2 dengan nilai silhoutte yang paling baik.



Gambar 7 Hasil sebaran data dengan 2 klaster



Gambar 8 Hasil sebaran data dengan 3 klaster



Gambar 9 Hasil sebaran data dengan 4 klaster

Dari hasil *benchmark* antara nilai *mean* masing-masing pengujian dengan 100 iterasi untuk KMeans *Clustering* dan KMeans PSO didapatkan hasil quantization error dari Kybrid KMeans PSO lebih kecil yaitu 2.920 dari pada Kmeans yaitu 2.939 berikut pada Tabel 3.

Tabel 3 Hasil Benchmark KMeans dan KMeans PSO

Metode	**sse**	**silhouette**	**calinski**	**davies**	**quantization**
K-Means	17.288	**0.700**	560.243	**0.643**	2.939
KMeans PSO	**17.255**	0.691	**562.001**	0.668	**2.920**

Dari penelitian yang dilakukan pada Tabel 3, terlihat ada perbandingan antara dua metode pengelompokan data, yaitu K-Means dan KMeans *Particle Swarm Optimization* (PSO). Hasil penelitian diukur menggunakan beberapa metrik evaluasi untuk mengukur kualitas pengelompokan data.

1. *Sum of Squared Errors* (SSE)

SSE mengukur sejauh mana setiap titik data dalam kelompoknya dari pusat kelompoknya. Semakin rendah nilai SSE, semakin padat dan dekat titik-titik dalam kelompok. Dari hasil, kedua metode memiliki nilai SSE yang hampir sama, namun KMeans PSO memiliki keunggulan dengan nilai yang lebih rendah (17.255 dibandingkan dengan 17.288 dari K-Means).

2. *Silhouette Score*

Skor silhouette mengukur seberapa baik setiap titik data berada dalam kelompoknya dibandingkan dengan kelompok lainnya. Rentang skor silhouette adalah -1 hingga 1, di mana nilai lebih tinggi menunjukkan pengelompokan yang lebih baik. Dalam penelitian ini, K-Means memiliki skor silhouette yang sedikit lebih tinggi dibandingkan dengan KMeans PSO (0.700 dibandingkan dengan 0.691).

3. *Calinski-Harabasz Index*

Indeks Calinski-Harabasz juga dikenal sebagai Variance Ratio Criterion. Ini mengukur rasio antara dispersi antara kelompok dan dispersi dalam kelompok. Nilai indeks yang lebih tinggi menunjukkan pengelompokan yang lebih baik. Dalam penelitian ini, KMeans PSO memiliki nilai indeks Calinski-Harabasz yang lebih tinggi dibandingkan dengan K-Means (562.001 dibandingkan dengan 560.243).

4. *Davies-Bouldin Index*

Indeks Davies-Bouldin mengukur seberapa baik setiap kelompok terpisah satu sama lain. Semakin rendah nilai indeks, semakin baik pengelompokan. Hasil menunjukkan bahwa Kmeans memiliki nilai indeks Davies-Bouldin yang sedikit lebih rendah dibandingkan dengan K-Means (0.643 dibandingkan dengan 0.668).

5. *Quantization Error*

Metrik Quantization Error adalah sebuah metode evaluasi yang digunakan untuk mengukur sejauh mana titik-titik data dalam dataset dapat direpresentasikan oleh pusat kelompok terdekat dalam algoritma pengelompokan. Metrik ini sering digunakan dalam konteks algoritma pengelompokan.

Secara lebih teknis, Quantization Error (QE) mengukur rata-rata pada jarak antara setiap titik data dan pusat kelompok yang menjadi *centroid* atau representasi dari kelompok tersebut. Semakin rendah nilai QE, semakin baik titik-titik data direpresentasikan oleh pusat kelompoknya, yang mengindikasikan pengelompokan yang lebih baik. Dalam hal ini, KMeans PSO memiliki nilai quantization yang lebih rendah dibandingkan dengan K-Means (2.920 dibandingkan dengan 2.939).

Berdasarkan hasil penelitian ini, meskipun ada perbedaan yang kecil, KMeans PSO memberikan kualitas pengelompokan yang lebih baik dibandingkan dengan metode K-Means, mengingat performa yang lebih baik dalam beberapa metrik evaluasi seperti *Sum of Squared Errors* (SSE), *Calinski-Harabasz Index*, dan nilai *quantization* yang lebih rendah.

4. KESIMPULAN DAN SARAN

Dari hasil penelitian ini, dapat disimpulkan diantaranya :
a) Algoritma KMeans dan PSO memberikan kualitas pengelompokan yang lebih baik jika dikomparasikan dengan metode K-Means, mengingat performa yang lebih baik dalam beberapa metrik evaluasi seperti Sum of Squared Errors (SSE) 17.255, Calinski-Harabasz Index 562.001, dan nilai quantization yang lebih rendah 2.920.
b) Pembagian data menghasilkan 2 klaster terbaik dengan sebaran yang masing -masing klaster 0 sebanyak 64 data dan klaster 1 sebanyak 277 data. Dimana klaster 0 adalah Penjualan paket data dengan transaksi tinggi namun pendapatan rendah (harga satuan yang rendah) fokus promosinya untuk menaikkan jumlah transaksi dengan penurunan harga secara paket/bundle program. Untuk klaster 1 adalah penjualan paket data dengan transaksi rendah namun pendapatan yang tinggi karena harga paket satuan tinggi/bersaing fokus promosinya untuk menaikkan harga satuan dan notifikasi untuk masing-masing region dengan transaksi yang rendah. Pembagian sebaran data berdasarkan region masing-masing klaster terbagi dalam klaster 0 region CENTRAL sebanyak 15 data, EAST 19 data, JABO 17 data dan WEST 13 data sedangkan dalam klaster 1 region CENTRAL sebanyak 72 data, EAST 87 data, JABO 83 data dan WEST 35 data.

Hasil klasterisasi dengan algoritma K-Means dan PSO dapat membantu dalam mengidentifikasi segmen pelanggan yang berbeda berdasarkan perilaku dan preferensi mereka terhadap paket data XL Axiata. Dengan pemahaman yang lebih baik tentang kelompok pelanggan yang berbeda, tim penjualan dan distribusi dapat merancang dan melakukan strategi promosi yang lebih tepat sasaran, menyediakan penawaran yang sesuai dengan kebutuhan dan preferensi pelanggan, serta meningkatkan efektivitas upaya penjualan.

Penelitian ini mengoptimasi klasterisasi dan memberikan data kluster terbaik untuk membantu strategi promosi penjualan paket data. Hasil penelitian ini dapat menjadi sumbangsih kontribusi yang positif bagi industri telekomunikasi dan menginspirasi penelitian lebih lanjut dalam penggunaan algoritma klasterisasi untuk pengembangan strategi promosi dan penjualan.

DAFTAR PUSTAKA

[1] M. Ikbal, S. Saputra, Y. A. Putri, and H. Nurhijmah, “Studi Komparasi Tingkat Ekonomis Aplikasi Myxl Dan Mytelkomsel Berbasis Mobile,” 2022.

[2] R. Dwi Apriansa, I. N. Farida, and U. Mahdiyah, “Sistem Rekomendasi Penentuan Poin Produk Menggunakan Algoritma FP-Growth Dan K-Means Clustering,” 2022.

[3] J. Kennedy and R. Eberhart, “Particle swarm optimization,” in *Proceedings of ICNN’95 - International Conference on Neural Networks*, 1995, pp. 1942–1948 vol.4. doi: 10.1109/ICNN.1995.488968.

[4] D. W. van der Merwe and A. P. Engelbrecht, “Data clustering using particle swarm optimization,” in *The 2003 Congress on Evolutionary Computation, 2003. CEC ’03.*, 2003, pp. 215-220 Vol.1. doi: 10.1109/CEC.2003.1299577.

[5] R. A. Indraputra and R. Fitriana, “K-Means Clustering Data COVID-19,” 2022.

[6] H. Syukron, M. Fauzi Fayyad, F. Junita Fauzan, Y. Ikhsani, and U. Rizkya Gurning, “MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison K-Means K-Medoids and Fuzzy C-Means for Clustering Customer Data with LRFM Model "Perbandingan K-Means K-Medoids dan Fuzzy C-Means untuk Pengelompokan Data Pelanggan dengan Model LRFM,” vol. 2, pp. 76–83, 2022.

[7] R. Adha, N. Nurhaliza, and U. Soleha, “Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia,” *Jurnal Sains, Teknologi dan Industri*, vol. 18, no. 2, pp. 206–211, 2021, [Online]. Available: https://covid19.who.int.

[8] I. G. Krisna *et al.*, “Perbandingan Pengelompokan Metode PSO K-Means Dan Tanpa PSO Dalam Pengelompokan Data Alert,” 2022, [Online]. Available: www.CTUMalware.com

[9] A. Rachwał *et al.*, "Determining the Quality of a Dataset in Clustering Terms," *Applied Sciences (Switzerland)*, vol. 13, no. 5, Mar. 2023, doi: 10.3390/app13052942.

[10] S. Handoko, F. Fauziah, and E. T. E. Handayani, "Implementasi Data Mining Untuk Menentukan Tingkat Penjualan Paket Data Telkomsel Menggunakan Metode K-means Clustering," *Jurnal Ilmiah Teknologi dan Rekayasa*, vol. 25, no. 1, pp. 76–88, 2020, doi: 10.35760/tr.2020.v25i1.2677.