SKRIPSI

IMPLEMENTASI *K-MEDOIDS* DAN *K-MEANS CLUSTERING* PADA DATA STEAM

Oleh :
Candra Wisantra
065116044

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PAKUAN
BOGOR
2023

SKRIPSI

IMPLEMENTASI *K-MEDOIDS* DAN *K-MEANS CLUSTERING* PADA DATA STEAM

Diajukan sebagai salah satu syarat untuk memperoleh
Gelar Sarjana Komputer Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam

Oleh :
Candra Wisantra
065116044

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PAKUAN
BOGOR
2023

HALAMAN PENGESAHAN

Judul : Implementasi *K-Medoids* dan *K-means Clustering* pada data *Steam*
Nama : Candra Wisantra
NPM : 065116044

Mengesahkan,

Pembimbing Pendamping FMIPA-UNPAK	Pembimbing Utama FMIPA-UNPAK
Dr. Fajar Delli W., SSI.MM., M.Kom.	Dr. Prihastuti Harsani, M.Si

Mengetahui,

Ketua Program Studi Ilmu Komputer FMIPA-UNPAK	Dekan FMIPA-UNPAK
Arie Qur'ania, M.Kom	Asep Denih, S.Kom., M.Sc., Ph.D.

PERNYATAAN KEASLIAN KARYA TULIS SKRIPSI

Dengan ini saya menyatakan bahwa:

Sejauh yang saya ketahui, karya tulis ini bukan merupakan karya tulis yang pernah dipublikasikan atau sudah pernah dipakai untuk mendapatkan gelar sarjana di Universitas lain, kecuali pada bagian-bagian dimana sumber informasinya dicantumkan dengan cara referensi yang semestinya.

Demikian pernyataan ini saya buat dengan sebenar-benarnya. Apabila kelak dikemudian hari terdapat gugatan, penulis bersedia dikenakan sanksi sesuai dengan peraturan yang berlaku.

Bogor, Juli 2023

Candra Wisantra

065116044

PERNYATAAN PELIMPAHAN SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

==========

Saya yang bertandatangan di bawah ini :

Nama	: Candra Wisantra
NPM	: 065116044
Judul Skripsi	: Implementasi *K-Medoids* dan *K-means Clustering* pada data *Steam*

Dengan ini saya menyatakan bahwa Paten dan Hak Cipta dari produk Skripsi dan Tugas Akhir di atas adalah benar karya saya dengan arahan dari komisi pembimbing dan belum diajukan dalam bentuk apapun kepada perguruan tinggi manapun.

Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan Paten, hak cipta dari karya tulis saya kepada Universitas Pakuan.

Bogor, Juli 2023

Candra Wisantra
065116044

RIWAYAT HIDUP

Penulis Bernama lengkap Candra Wisantra, dilahirkan di Bogor pada tanggal 3 Maret 1995, dari Ayah yang bernama Rosa Wisantra dan Ibu yang bernama Syamsiah, penulis adalah anak kedua dari 4 bersaudara kaka pertama bernama Mulya Warman, adik pertama yang bernama Aura Amalia dan adik kedua yang bernama Bayu Wisantra.

Penulis mengawali Pendidikan di Sekolah Dasar Negri Pamoyanan 3 dan lulus pada tahun 2007. Pada tahun 2010 penulis menamatkan Pendidikan Sekolah Menengah Pertama di SMPN 13 Kota Bogor. Penulis melanjutkan Pendidikan di SMK Negri 4 Kota Bogor dan menamatkan Pendidikan pada tahun 2013.

Pada tahun 2016 penulis kemudian meneruskan Pendidikan di Universitas Pakuan Bogor, Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, selama di Universitas Pakuan, penulis melakukan praktek lapang di SMA Ananda kota Bogor, untuk menyelesaikan tugas praktek lapang dengan merancang system informasi berbasis android. Pada tahun 2023, penulis menyelesaikan penelitian dengan judul "Implementasi *K-Medoids* Dan *K-Means Clustering* pada Data *Steam*".

RINGKASAN

Candra Wisantra, 2020. Implementasi *K-Medoids* dan *K-means Clustering* pada data *Steam*. Dibawah bimbingan Dr. Prihastuti Harsani, M.Si. dan Dr. Fajar Delli W.,SSI.MM.,M.Kom.

Penelitian ini melakukan analisis pada data Steam menggunakan metode *K-Means clustering* dan *K-Medoids*, Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, Data yang digunakan dalam penelitian ini berupa data informasi game digital pada platform *Steam*, data Steam ini berisikan 27.075 baris data dan 18 kolom informasi Tujuan dari penelitian ini adalah menerapkan metode *K-Means clustering* dan *K Medoids* untuk menganalisis data Steam serta menganalisis perbedaan antara kedua metode tersebut.

Data Steam akan masuk ke dalam proses data mining, dalam proses data mining terdiri dari beberapa tahapan seperti preprocessing data, transformation data tujuan dari kedua proses tersebut untuk mempersiapkan data untuk siap diolah menggunakan metode *K-Means clustering* dan k- medoids, setelah metode *K-Means clustering* dan *K-Medoids* telah selesai maka masuk ke proses evaluasi untuk membandingkan metode *K-Means* dan *K-Medoids*, metode yang digunakan untuk mengevaluasi yaitu menggunakan metode *DBI (Davies-Bouldin Index)*, setelah evaluasi selesai maka akan masuk ke proses knowledge untuk menarik informasi baru yang diperoleh dari hasil *clustering* dari metode k means maupun k medoids serta perbedaan dari kedua metode *clustering* tersebut. Informasi yang didapatkan yaitu menunjukkan bahwa *K-Means clustering* lebih unggul dalam mengolah data dari platform Steam daripada *K-Medoids*. Hal ini didapatkan dengan cara membandingkan nilai evaluasi dari kedua metode *clustering* tersebut *K-Means* memiliki nilai 0,4644 sedangkan K Medoids memiliki nilai 1,5604, semakin kecil nilai dbi maka semakin bagus hasilnya, selain itu hasil *clustering* menggunakan K-Means *clustering* lebih beragam dibandingkan menggunakan *K-Medoids* yang menunjukan *K-Means clustering* lebih unggul.

KATA PENGANTAR

Puji syukur kehadirat Allah SWT, karena rahmat dan hidayah-Nya penulis dapat menyelesaikan proposal penelitian ini yang berjudul "Implementasi *K-Medoids* dan *K-means Clustering* pada data *Steam*" Penulisan Hasil Penelitian ini merupakan salah satu syarat kelulusan di Program Studi Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Pakuan Bogor. Dalam kesempatan ini, penulis tidak lupa meyampaikan terima kasih kepada semua pihak yang telah membantu kelancaran penyusunan Skripsi ini, khususnya kepada:

1. Dr. Prihastuti Harsani, M.Si. selaku Dosen Pembimbing utama yang senantiasa memberikan pengarahan dan bimbingan selama menyusun tugas akhir ini.
2. Dr. Fajar Delli W.,SSI.MM.,M.Kom, selaku Dosen Pembimbing pendamping yang senantiasa memberikan pengarahan dan bimbingan selama penyusunan tugas akhir ini.
3. Arie Qur'ania, M.Kom selaku Ketua Program Studi Ilmu Komputer di FMIPA – Universitas Pakuan.
4. Kedua Orangtua yang sudah mendukung, mendoakan dan memberikan motivasi demi terselesaikannya Skripsi ini.
5. Serta rekan-rekan mahasiswa yang telah memberikan dorongan dan motivasinya.

Menyadari bahwa Skripsi ini masih jauh dari kesempurnaan, karena keterbatasannya pengetahuan serta kemampuan yang dimiliki. Oleh sebab itu kritik dan saran sangat diharapkan yang bersifat untuk membangun demi penyempurnaan penyusunan skripsi ini agar bisa lebih baik.

Bogor, Juli 2023

Candra Wisantra

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
LATAR BELAKANG

1.1 Latar Belakang

Pada era digital saat ini, jumlah data yang dihasilkan dari berbagai sumber seperti media sosial, situs web seperti *e-commerence*, data industri mengenai *tren* pasar dan sensor *IoT* semakin meningkat dengan cepat. Data ini bisa menjadi informasi berharga yang dapat digunakan untuk mengambil keputusan yang lebih baik dalam berbagai bidang seperti bisnis, ilmu pengetahuan, dan teknologi. Namun, untuk mengambil manfaat dari data ini, diperlukan alat dan teknik untuk menganalisis dan mengelompokkan data menjadi kelompok yang signifikan.

Metode *clustering* atau pengelompokan merupakan salah satu teknik dalam menganalisis data yang digunakan untuk mengelompokkan data menjadi beberapa kelompok berdasarkan kesamaan karakteristik. Salah satu metode *clustering* yang populer digunakan adalah *k-means* dan *k-medoids*. Kedua metode ini memiliki keunggulan dan kelemahannya masing-masing, dan telah digunakan dalam berbagai bidang seperti ilmu sosial, keuangan, kesehatan dan lain – lain.

Namun, perbedaan antara k-means dan k-medoids belum sepenuhnya dimengerti. Beberapa penelitian telah dilakukan untuk membandingkan kedua metode, dalam penerapannya, pemilihan antara metode k-means dan k-medoids seringkali bergantung pada konteks dan karakteristik data yang spesifik. Oleh karena itu, tujuan dari penelitian ini adalah untuk membandingkan performa *k-means* dan *k-medoids* dalam mengelompokkan data berdasarkan karakteristik dari data yang akan digunakan.

Dalam penelitian ini, peneliti akan menggunakan data Steam yang didapat dari Kaggle dan menerapkan k-means dan k-medoids pada data tersebut. Sedikit pengenalan mengenai Steam ialah distribusi digital dan toko online yang dikelola oleh *Valve Corporation*, sebuah perusahaan pengembang game terkenal, Platform ini didedikasikan untuk game, dan menyediakan ribuan game yang dapat diunduh dan dimainkan pada PC, serta beberapa konsol game. Hasil penelitian kami diharapkan dapat membantu pemilihan metode *clustering* yang tepat untuk data dengan karakteristik tertentu.

Adapun penelitian sebelumnya oleh (Ahuja et al., 2019)"*movie recommender system using K-means clustering and K-Nearest neghbor*" penulis menggunakan data rating film dari situs IMDB dan membangun model rekomendasi film dengan menggunakan *K-Means* untuk mengelompokkan pengguna berdasarkan preferensi film mereka. Setelah itu, penulis menggunakan algoritma *KNN* untuk merekomendasikan film kepada pengguna berdasarkan kelompok pengguna yang memiliki preferensi film yang sama. Kesimpulan dari penggunaan teknik clustering *K-Means* dan algoritma *KNN* dalam membangun sistem rekomendasi film, model yang dikembangkan berhasil memberikan rekomendasi film yang akurat dan dapat membantu pengguna untuk menemukan film yang sesuai dengan preferensi mereka.

Adapun penelitian sebelumnya oleh (Rao et al., 2019)"*Interval data-based k-means clustering method for traffic state identification at urban intersections*" membahas tentang pengembangan metode *clustering k-means* berbasis data interval untuk mengidentifikasi keadaan lalu lintas di persimpangan perkotaan. Penulis menjelaskan prinsip dasar dari *k-means clustering* dan interval data, dan kemudian mengajukan metode baru yang memanfaatkan data interval pada penghitungan jarak

antar titik dan centroid. Hasil eksperimen menunjukkan bahwa metode yang diusulkan berhasil meningkatkan akurasi dalam mengidentifikasi keadaan lalu lintas di persimpangan perkotaan, dibandingkan dengan metode clustering lainnya seperti *k-medoids* dan *fuzzy c-means*. Kesimpulannya, jurnal ini menyajikan pengembangan metode clustering k-means berbasis data interval untuk mengidentifikasi keadaan lalu lintas di persimpangan

Adapun penelitian sebelumnya oleh (Luthfi et al., 2021)"Analisis perbandingan metode *hirearchical, k-means,* dan *k-medoids clustering* dalam pengelompokkan indeks pembangunan manusia Indonesia" Penulis menjelaskan prinsip dasar dari kedua metode clustering yaitu hierarchical clustering, k-means *clustering*, dan *k-medoids clustering* dan kemudian melakukan eksperimen pada data indeks pembangunan manusia Indonesia menggunakan kedua metode clustering tersebut. Hasil eksperimen menunjukkan bahwa metode *k-medoids* clustering menghasilkan pengelompokkan yang lebih baik daripada metode *hierarchical* dan *k-means clustering*, dilihat dari nilai indeks validitas internal dan eksternal. Selain itu, metode k-medoids clustering juga mampu menghasilkan representasi pusat kelompok yang lebih baik daripada metode lainnya. Kesimpulannya, jurnal ini menyajikan perbandingan antara dua metode clustering, yaitu *hierarchical, k-means, dan k-medoids clustering* dalam pengelompokkan *indeks* pembangunan manusia di Indonesia. Metode *k-medoids clustering* menghasilkan pengelompokkan yang lebih baik dan lebih representatif daripada metode lainnya.

1.2 Tujuan

Tujuan dari penelitian ini adalah menerapkan metode *K-means clustering* dan *K-medoids* untuk menganalisis data Steam serta menganalisis perbedaan antara kedua metode tersebut.

1.3 Ruang Lingkup

Untuk lebih terarah perlunya dibuat batasan masalah, adapun ruang lingkup penelitian ini meliputi :

1. Studi kasus hanya berdasarkan data Steam berasal dari kaggle dengan data diambil pada Mei 2019 (https://www.kaggle.com/datasets/nikdavis/steam-store-games)
2. Metode yang digunakan yaitu *K-means clustering* dan K-medoids
3. Untuk menentukan jumlah cluster menggunakan metode elbow & silhouette diagram
4. Tujuan hanya memberikan gambaran umum data deskriptif melalui perbandingan.

1.4 Manfaat

Manfaat penelitian ini diantaranya :

1. Diharapkan bisa menambah wawasan dalam menganalisis data
2. Diharapkan dapat berkontribusi dalam bidang data mining
3. Diharapkan dalam hasil penelitian untuk menemukan fenomena atau keadaan tertentu yang terdapat pada data Steam.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

2.1.1 Clustering

Clustering dapat dianggap sebagai masalah pembelajaran tanpa pengawasan, jadi setiap masalah semacam ini berkaitan dengan menemukan struktur dalam kumpulan data yang tidak berlabel, oleh karena itu *cluster* merupakan kumpulan objek yang mirip di antara kumpulan yang tidak mirip dengan objek milik *cluster* yang lain.(Soni Madhulatha, 2012). *Clustering* merupakan pusat dari banyak penelitian *bioinformatika* berbasis data dan menyajikan metode komputasi yang kuat, secara khusus *clustering* membantu menganalisis data yang tidak terstruktur dan berdimensi tinggi dalam bentuk urutan, ekspresi, teks dan gambar (Karim et al., 2021).

2.1.2 K-Medoids

Algoritma *K-medoids* biasa juga disebut algoritma *PAM(Partitioning Around Medoids)* merupakan algoritma *representative cluster* yaitu *medoid*, algoritma *K-medoid* menggunakan objek sebagai perwakilan pusat cluster untuk tiap cluster.(Kamila & Khairunnisa, 2019) Algoritma *K-medoid* lebih baik daripada metode *K-means* dalam menangani noise dan *outlier*, karena metodenya tidak terlalu dipengaruhi oleh outlier atau nilai ekstrim lainnya (Rahman et al., 2020)

Tahapan *K Medoids* :

1. Menentukan jumlah cluster
2. Mengambil data secara *random* untuk dijadikan centroid awal
3. Menghitung tiap data ke centroid terdekat
4. Menentukan titik centroid
5. Mengelompokan data

2.1.3 K-Means Clustering

Algoritma *K-means* merupakan salah satu dari sepuluh besar algoritma untuk memecahkan masalah pengelompokan, *K-means clustering* algoritma yang didasarkan pada perhitungan fungsi jarak, algoritma berlangsung dalam dua fase, penugasan fase yang menetapkan setiap titik *cluster* berdasarkan jarak titik terdekat dan fase pembaruan untuk menghitung ulang *centroid*, setiap *cluster* berdasarkan tiap titik, setiap fase diulang secara berurutan sampai centroid berhenti bergerak (Xia et al., 2022) Algoritma *K-means* memiliki banyak keunggulan seperti ide matematika yang sederhana, konversi yang cepat dan implementasi yang mudah, oleh karena itu bidang pengaplikasian sangat luas termasuk sebagai jenis klasifikasi dokumen, musik, film, klasifikasi perilaku pembelian, sistem rekomendasi berdasarkan minat pengguna dan sebagainya (Chunhui Yuan, 2019)

Tahapan *K-means clustering* :

1. Menentukan jumlah cluster
2. Mengambil data secara random untuk dijadikan centroid awal
3. Menghitung tiap data ke centroid terdekat

4. Menentukan titik centroid
5. Mengelompokan data

2.1.4 Data Mining

Data mining merupakan analisis dari pengamatan sekumpulan data untuk mencari hubungan yang tidak terduga dan menyimpulkan data dengan beberapa cara sehingga data tersebut memiliki nilai. Data mining merupakan kumpulan dari beberapa bidang keilmuan yang menyatukan Teknik dari statistic, database, visualisasi, *machine learning* untuk penyelesaian permasalahan pengambilan informasi dari database(Utomo & Mesran, 2020). Data mining merupakan proses pencarian corak atau informasi yang bernilai dalam kumpulan data dengan menggunakan metode atau Teknik tertentu (Putra & Wadisman, 2018).

2.1.5 Python

Python merupakan bahasa pemrograman berorientasi objek, ditafsirkan, dan interaktif. Ini menyediakan struktur data tingkat tinggi seperti daftar, tupel, set, array asosiatif (disebut kamus), dinamis mengetik dan mengikat, modul, kelas, pengecualian, manajemen memori otomatis, dll. Ini juga digunakan untuk sistem komputasi paralel dan memiliki sintaks yang relatif sederhana dan mudah untuk pengkodean dan tetap saja itu adalah bahasa pemrograman yang kuat. *Python* memiliki juru bahasa untuk java yang dikenal sebagai Python, yang mirip dengan juru bahasa untuk bahasa C. *Python* memiliki banyak keunggulan dibandingkan apapun bahasa lain, seperti itu memiliki *varietas* perpustakaan yang mengurangi kode menjadi sepertiga untuk programmer dan karena ini Python telah mencapai puncak tertinggi dalam hal Pembelajaran Mesin (Dhruv et al., 2021).

2.1.6 Davies Bouldin Index

Davies Bouldin Index (DBI) merupakan salah satu metode yang digunakan untuk mengukur validitas *cluster* pada suatu metode *clustering*, pengukuran dengan DBI dapat memaksimalkan jarak antar *cluster* serta pada saat yang sama mencoba meminimalkan jarak antar titik dalam sebuah cluster. Jika jarak antar cluster maksimum, berarti kesamaan karakteristik tiap cluster tidak banyak menyebabkan perbedaan antar cluster terlihat lebih jelas. Apabila sebaliknya jika jarak *intra-cluster* minimal kemungkinan setiap objek dalam cluster memiliki tingkat kemiripan karakter yang tinggi. Hasil dari cluster diperoleh dari usulan penentuan centroid kemudian dievaluasi dengan metode *DBI.* Sehingga dapat diketahui korelasi dari penentuan metode centroid berdasarkan *sum of squared error* terhadap peningkatan kualitas cluster berdasarkan nilai dbi yang diperoleh (Jumadi Dehotman Sitompul et al., 2019).

2.1.7 Metode *Elbow*

Metode *Elbow* merupakan suatu metode yang bertujuan untuk menentukan jumlah cluster terbaik melalui perbandingan persentase antar jumlah *cluster* yang akan membentuk grafik (Putu et al., 2018), dari grafik yang dihasilkan dapat menentukan jumlah cluster yang terbaik, biasanya grafik yang dihasilkan akan membentuk seperti pola siku oleh karena itu metode ini mempunyai nama *Elbow*.

2.1.8 Google Colab

Google colab merupakan produk analisis Google yang memungkinkan siapa saja melalui browser untuk menulis dan menjalankan kode *Python*, serta sangat cocok untuk pembelajaran komputer, pemrosesan data, dan pendidikan(Velu & Whig, 2021). Google Colab tidak memerlukan konfigurasi serta menyediakan akses ke GPU dan TPU gratis,Google Colab merupakan bentuk modifikasi dari Jupyter Notebook yang dapat digunakan untuk *machine learning ataupun deep learning* (Febrian Sengkey et al., 2020).

2.1.9 Silhouette Diagram

Silhouette diagram merupakan metode yang dimanfaatkan untuk menentukan jumlah *cluster* yang optimal dalam data, dengan mengukur kemiripan objek dengan cluster sendiri dibandingkan dengan cluster lainnya (Dewi & Pramita, 2019).

2.1.10 Steam

Steam merupakan sebuah platform distribusi digital yang dimiliki oleh perusahaan game *Valve Corporation*, Steam juga menyediakan fitur untuk berinteraksi dengan teman-teman, memainkan game bersama dan partisipasi dalam komunitas. *Steam* merupakan platform game dan situs jejaring sosial terkemuka, yang memungkinkan penggunanya untuk membeli dan menyimpan game, selain itu *Steam* menyediakan database besar informasi seputar game, pemain serta perilaku game (Balmford, 2020).Unduhan game digital telah menjadi bentuk paling umum dari pembelian video game pc diseluruh dunia, kekuatan pasar utama dalam gerakan distribusi digital ini adalah platform *Steam, Steam* mendominasi pasar video game sebesar 75% dari 92% game yang dijual secara digital (De Luisa et al., 2021).

2.2 Penelitian Terdahulu

Penulis menggunakan penelitian terdahulu sebagai acuan melakukan penelitian, dari penelitian terdahulu penulis tidak menemukan penelitian dengan judul yang sama, namun penulis menggunakan beberapa penelitian sebagai referensi sebagai berikut :

1. Nama : Ahuja et al., 2019
   Judul : *Movie Recommender System Using K-Means Clustering AND K- Nearest Neighbor*
   Isi : Metode *K-Means Clustering* digunakan untuk mengelompokkan pengguna berdasarkan preferensi film mereka, sementara metode *KNN* digunakan untuk merekomendasikan film kepada pengguna berdasarkan preferensi film mereka dan preferensi pengguna dengan preferensi serupa. Penulis menggunakan data dari situs web IMDb dan membaginya menjadi tiga set data: latihan, validasi, dan pengujian. Mereka mengevaluasi kinerja sistem rekomendasi mereka menggunakan metrik seperti presisi, recall, dan F1-score. Hasil eksperimen menunjukkan bahwa sistem rekomendasi yang diusulkan menghasilkan kinerja yang baik, dengan presisi mencapai 0,70 dan recall mencapai 0,74 pada set pengujian. Penulis juga mencatat bahwa metode *K-Means Clustering* dan *KNN* bekerja dengan baik dalam sistem rekomendasi film. Penulis berharap bahwa hasil penelitian

mereka dapat membantu pengembangan sistem rekomendasi film yang lebih akurat dan efektif di masa depan.

2. Nama : Rao et al., 2019
   Judul : *Interval data-based k-means clustering method for traffic state identification at urban intersections*
   Isi : Jurnal ini mengusulkan metode baru untuk mengidentifikasi keadaan lalu lintas di persimpangan perkotaan menggunakan metode *clustering k-means* berbasis interval data. Metode ini melibatkan pengumpulan data aliran lalu lintas dari sensor yang dipasang di persimpangan, dan kemudian melakukan preprocessing data untuk mendapatkan interval data yang merepresentasikan keadaan lalu lintas. Selanjutnya, algoritma clustering k-means diterapkan pada interval data untuk mengidentifikasi berbagai keadaan lalu lintas. Penulis melakukan eksperimen pada data lalu lintas dunia nyata yang dikumpulkan dari dua persimpangan di China, dan membandingkan performa metode mereka dengan metode clustering lainnya seperti metode *k-medoids* dan *fuzzy*. Hasilnya menunjukkan bahwa metode yang diusulkan memiliki kinerja yang lebih baik dibandingkan dengan metode clustering lainnya dalam hal akurasi dan efisiensi, dan dapat secara efektif mengidentifikasi keadaan lalu lintas yang berbeda, seperti lalu lintas lancar, kemacetan, dan antrian. Penulis menyimpulkan bahwa metode K-means dapat berguna untuk manajemen dan pengendalian lalu lintas di persimpangan perkotaan. Secara keseluruhan, jurnal ini mempresentasikan metode baru untuk mengidentifikasi keadaan lalu lintas di persimpangan perkotaan, yang dapat membantu meningkatkan manajemen lalu lintas dan mengurangi kemacetan
3. Nama : Luthfi et al., 2021
   Judul : Analisis perbandingan metode *hirearchical, k-means,* dan *k-medoids clustering* dalam pengelompokkan indeks pembangunan manusia Indonesia
   Isi : membahas tentang perbandingan tiga metode *clustering*, yaitu *hierarchical clustering, k-means clustering*, dan *k-medoids clustering* untuk mengelompokkan data indeks pembangunan manusia Indonesia. Penelitian ini bertujuan untuk mengetahui metode clustering mana yang paling efektif dan efisien dalam mengelompokkan data indeks pembangunan manusia. Hasil penelitian menunjukkan bahwa metode k-means clustering memiliki tingkat efektivitas yang lebih tinggi dibandingkan dengan metode hierarchical clustering dan k-medoids clustering. Selain itu, metode k-means clustering juga lebih efisien dalam hal waktu komputasi dibandingkan dengan metode hierarchical clustering. Namun, metode k-medoids clustering memiliki keunggulan dalam hal robustness dan kehandalan dalam mengatasi noise atau data pencilan. Oleh karena itu, penelitian ini merekomendasikan penggunaan metode k-means clustering dalam pengelompokkan data indeks pembangunan manusia Indonesia dengan mempertimbangkan faktor efektivitas dan efisiensi waktu komputasi.

2.3 Tabel Penelitian Terdahulu

Tabel perbandingan berfungsi untuk membandingkan penelitian terdahulu yang kita gunakan sebagai bahan acuan. Berdasarkan pembahasan pada penelitian terdahulu, dapat ditarik kesimpulan dan dimasukan kedalam tabel, acuan tabel dapat dilihat pada tabel 1

Tabel 1 Tabel Perbandingan Penelitian Terdahulu

No	Penelitian Terdahulu	Implementasi							
		Metode						Aplikasi	
		A	B	C	D	E	F	G	H
1	Ahuja et al., 2019	√					√	√	
2	Rao et al., 2019	√	√	√	√				
3	Luthfi et al., 2021	√	√			√			√
4	Candra Wisantra	√						√	

Keterangan :

A. *K-means clustering*
B. *K-medoids clustering*
C. *Fuzzy C-means clustering*
D. *Gustafson-Kessel clustering*
E. *Hierarchical Clustering*
F. *K-Nearest Neighbor*
G. *Python*
H. *R*

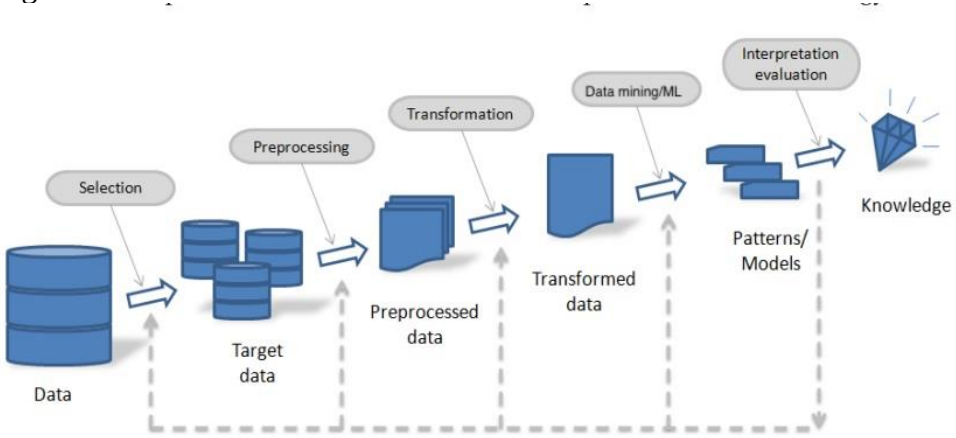
BAB III
METODE PENELITIAN

3.1 Metode Penelitian

Metode penelitian ini menjelaskan tentang metode penelitian yang digunakan dalam studi ini menganalisis data pada data Steam menggunakan teknik data mining, untuk penelitian ini menggunakan tahapan data mining atau *Knowledge Discovery in Database (KDD)*. Pada penelitian ini, digunakan teknik clustering untuk mengelompokkan game – game pada data Steam berdasarkan atribut-atribut yang dimilikinya. Metode perbandingan K-means dan K-medoids dipilih untuk digunakan dalam penelitian ini karena memiliki kinerja yang cukup baik dalam clustering data numerik, serta cocok untuk digunakan pada data yang memiliki ukuran yang cukup besar.

Pada penelitian ini, data yang digunakan adalah data *Steam* yang berisikan informasi mengenai game-game yang ada pada platforms *Steam*. Data ini terdiri dari 18 atribut seperti jumlah pemain, harga dan rating. Data steam ini didapatkan dari Kaggle, data yang terkumpul kemudian diproses menggunakan layanan cloud computing dari Google yaitu Google Colab dengan menggunakan bahasa pemrograman Python dan menggunakan metode *K-means* dan *K-medoids*

pada analisis clustering dengan metode *K-means* dan *K-medoids*, jumlah cluster yang optimal diperoleh dengan menggunakan metode *elbow*, setelah mendapatkan hasil dari metode *elbow* maka akan diverifikasi dengan metode silhouette diagram apakah benar merupakan cluster optimal. Setelah mendapatkan jumlah cluster optimal maka akan masuk ke tahap utama yaitu pengklasteran menggunakan metode k-means dan k-medoids, hasil dari analisis clustering tersebut kemudian dibandingkan untuk mengetahui perbedaan antara metode, untuk mengetahui kualitas clustering dengan menggunakan metode *DBI (Davies-Bouldin Index)* dalam metode ini nilai *DBI* yang lebih kecil menunjukan kualitas clustering yang yang lebih baik.



Gambar 1 Tahapan Data mining(Palacios et al., 2021)

3.1.1 Data Selection

Tahap *data selection* dalam proses data mining merupakan tahap awal yang sangat penting karena menentukan kualitas data yang akan digunakan dalam proses analisis selanjutnya. Pada tahap ini, peneliti harus memilih data yang relevan dan cocok dengan tujuan penelitian yang ingin dicapai. Dalam penelitian ini, data *Steam* dipilih sebagai sumber data karena berisikan informasi mengenai game-game pada platform Steam yang berkaitan dengan tujuan penelitian. Data Steam diperoleh melalui Kaggle, salah satu platform sumber data yang cukup populer dan terpercaya. Dalam memilih sumber data, penting untuk memperhatikan kepercayaan dan kredibilitas sumber data tersebut agar hasil analisis yang didapatkan dapat dipercaya.

3.1.2 Preprocessing Data

Preprocessing data merupakan proses untuk mempersiapkan serta membersihkan data sebelum dilakukannya tahap analisis, tahapan preprocessing sangat penting dilakukan untuk mencegah hasil yang buruk seperti keakuratan dari hasil analisis yang tidak bagus. Dalam tahapan preprocessing terjadi proses cleaning data dan data selection, pada proses cleaning data atau pembersihan data merupakan proses pembersihan data dari data yang hilang serta membersihkan duplikasi data, selanjutnya yaitu pemilihan variabel menggunakan metode korelasi pada tahap ini mengukur hubungan antara dua variabel, tujuan dari mengukur variabel ini untuk mengetahui apakah dua variabel ini memiliki korelasi positif atau negatif serta berapa kuat nilai korelasinya. Setelah melakukan pemilihan variabel data maka tahap selanjutnya yaitu pengecekan outlier menggunakan boxplot dan *IQR (Interquartile Range*) bertujuan untuk menemukan nilai nilai yang menyimpang yang mungkin dapat mempengaruhi hasil analisis data.

3.1.3 Transformation Data

Transformation data merupakan proses mengubah data dari format awal dirubah menjadi format yang dapat memudahkan proses algoritma yang akan digunakan, teknik yang akan digunakan pada tahap ini yaitu scaling data yaitu mengubah skala data supaya memiliki rentang nilai yang tidak terlalu jauh. Tujuan dari mengubah skala data untuk meningkatkan kualitas clustering.

3.1.4 Data Mining

Tahap data mining dapat dilakukan setelah data mengalami beberapa proses perubahan seperti *cleaning data* dan *transformation data* supaya data lebih siap untuk diproses. Pada tahap data mining data yang telah siap akan masuk ke tahap modeling yaitu merupakan tahapan membangun model atau algoritma yang dapat memprediksi atau mengambil informasi baru dari data. Dalam penelitian ini teknik pemodelan yang digunakan yaitu clustering, teknik clustering yang digunakan yaitu *k-means* dan *k-medoids,* untuk menentukan jumlah *cluster* optimal menggunakan metode *elbow* dan silhouette diagram.

3.1.5 Evaluation

Setelah melakukan *clustering* perlu dilakukannya evaluasi model untuk mengetahui kualitas hasil clustering. Dalam penelitian ini untuk mengevaluasi hasil

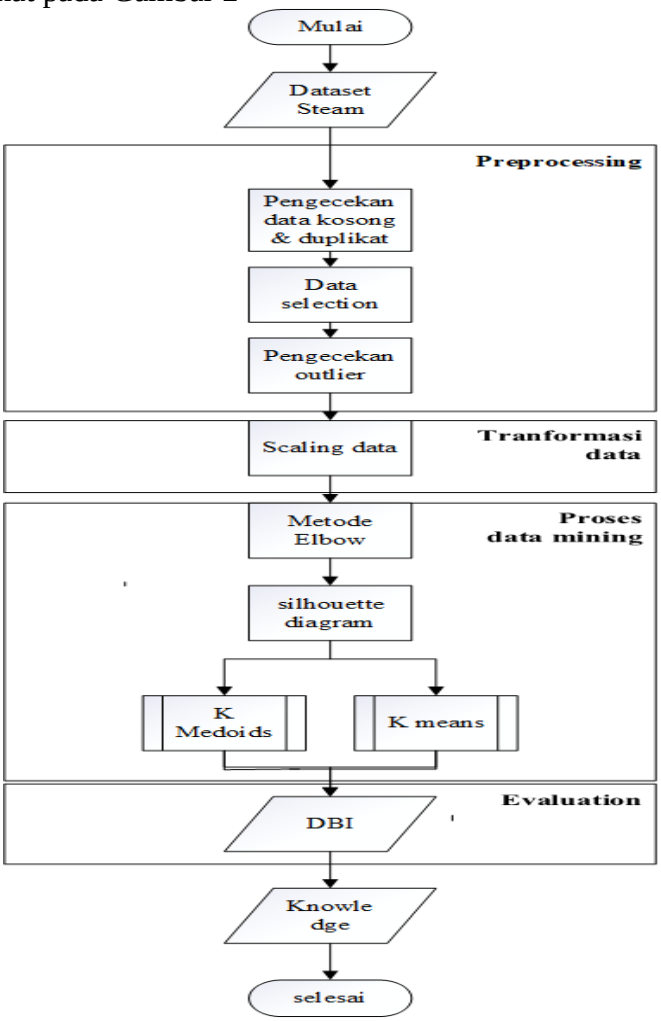
clustering menggunakan metode *DBI (Davies Bouldin Index)*, *DBI* digunakan untuk mengukur seberapa baik sebuah clustering, semakin rendah nilai *DBI* semakin baik clustering tersebut.

3.1.6 Knowledge

Knowledge merupakan informasi yang dihasilkan dari proses data mining, knowledge ini biasanya berupa pola atau tren yang ditemukan dari data yang telah diproses dan dianalisis. Dalam penelitian ini hasil clustering menggunakan K-means dan K-medoids dapat menghasilkan knowledge berupa kelompok data yang memiliki karakteristik atau pola tertentu,

3.2 Flowchart System

Flowchart system merupakan bagan yang akan menunjukan alur kerja yang akan menunjukan yang sedang dikerjakan di dalam sistem secara keseluruhan dan menjelaskan urutan dari prosedur prosedur yang ada dalam sistem. untuk flowchart sistem dapat dilihat pada Gambar 2



Gambar 2 Flowchart System

3.3 Alat dan Bahan

3.3.1 Alat

Alat yang digunakan dalam penelitian ini menggunakan spesifikasi sebagai berikut :

Software : Os Windows 10, Chrome
Hardware : Processor intel core i3, Hardisk 500gb, ram 10GB

3.3.2 Bahan

Bahan yang dibutuhkan dalam penelitian ini adalah jurnal terkait, data Steam yang diunduh dari situs web Kaggle yang bersifat open source dapat diakses secara public dengan format data csv.

BAB IV
PERANCANGAN DAN IMPLEMENTASI

4.1 Tahap Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle, yang merupakan platform penyedia data terbesar saat ini. Data yang digunakan dalam penelitian ini berupa data informasi game digital pada platform Steam, yang tersedia di link berikut : https://www.kaggle.com/datasets/nikdavis/steam-store-games. Data ini terdiri dari 27.075 baris data dan 18 kolom informasi seperti game, developer, rating dan harga.

4.2 Preprocessing Data

Pada tahapan ini dilakukan *preprocessing* data pada data Steam menggunakan bahasa pemrograman Python dan Google colab. Berikut ini merupakan langkah-langkah yang dilakukan dalam preprocessing data :

1. Import library yang dibutuhkan, seperti Pandas dan NumPy.
2. Data Steam yang akan digunakan dalam penelitian ini diunduh dari Kaggle dan dimuat menggunakan fungsi Pandas "read_excel()"
3. Melakukan pengecekan terhadap data, seperti cek jumlah baris dan kolom, cek nilai null pada setiap kolom, dan mengecek jenis data pada setiap kolom
4. Melakukan pengecekan data kosong dan menghapus baris tersebut apabila ditemukan data kosong
5. Melakukan pengecekan data duplikat dan menghapusnya apabila ditemukan data duplikat
6. Memilih variabel yang akan digunakan menggunakan korelasi
7. Pengecekan data outlier menggunakan boxplot.

4.3 Data Transformation

Setelah proses *preprocessing* data telah selesai dilakukan, selanjutnya dilakukan tahap *transformation* data untuk mengubah data Steam kedalam bentuk yang lebih siap digunakan pada tahap modeling. Salah satu teknik yang digunakan pada tahap ini adalah teknik scaling dengan metode standard scaler. Modul yang digunakan yaitu standardscaler dari library scikit-learn untuk melakukan scaling dengan memanggil fungsi "*fit_transform()*". Hasil dari transformasi akan disimpan pada variabel "x_train".

4.4 Data Mining

Setelah proses *data transformation* selanjutnya masuk ke tahap proses data mining yang bertujuan untuk mencari informasi baru yang terkandung dalam data yang biasanya berupa pola-pola tertentu. Untuk proses data mining menggunakan metode clustering menggunakan metode *k-means* dan *K medoids*, untuk menentukan jumlah cluster menggunakan metode *elbow* dan silhouette diagram.

4.4.1 Metode Elbow

Dalam penelitian ini, penentuan jumlah *cluster* yang optimal dilakukan menggunakan metode *elbow*. Metode *elbow* merupakan salah satu teknik yang cukup umum digunakan untuk menentukan jumlah cluster. Nilai inertia tiap *cluster* akan dibandingkan untuk menentukan jumlah cluster yang optimal, untuk menghitung nilai inersia menggunakan *library sklearn*. Untuk mempermudah dalam membandingkan nilai inertia ditampilkan menggunakan grafik yang memiliki lengkung seperti siku.

4.4.2 Silhouette Diagram

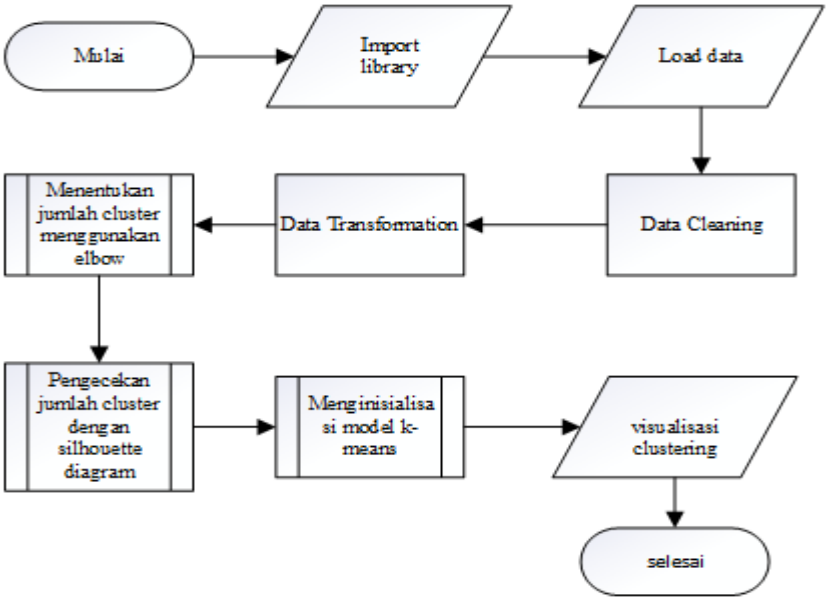
Langkah sebelumnya pada metode *elbow* didapatkan hasil cluster optimal yaitu 3 cluster, untuk membuktikan bahwa 3 cluster merupakan cluster optimal maka dilakukan pengecekan menggunakan silhouette diagram, fungsi *silhouette diagram* sama dengan metode *elbow* yaitu merupakan metode untuk menentukan jumlah cluster optimal.

4.4.3 K-Means Clustering

Setelah melalui berbagai tahapan preprocessing data, data transformation serta penentuan jumlah cluster menggunakan metode *elbow* dan silhouette diagram maka selanjutnya yaitu ke tahap clustering menggunakan metode K-means. Berikut ini merupakan tahapan perancangan dan implementasi k-means clustering :

1. Import library yang akan digunakan, seperti Pandas, Numpy dan Sklearn untuk menjalankan algoritma k means clustering.
2. Load data yang akan dianalisis
3. Melakukan tahapan preprocessing data.
4. Melakukan transformasi data
5. Menentukan jumlah cluster menggunakan metode elbow dan silhouette diagram
6. Inisialisasi model k means dengan jumlah cluster yang telah ditentukan
7. Visualisasi hasil k means clustering.

Berikut ini merupakan Flowchart dari k means clustering yang ditunjukan pada Gambar 3



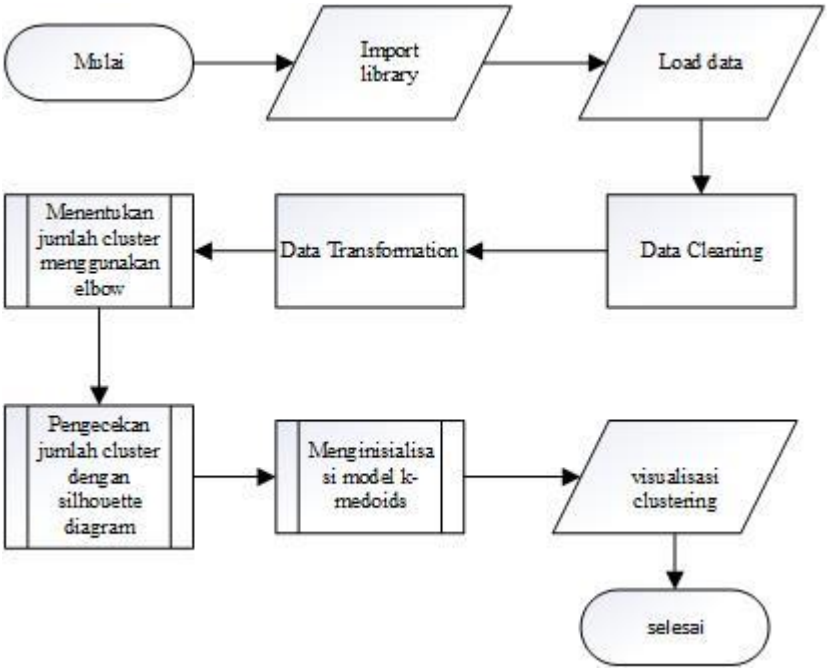
Gambar 3 Flowchart K-means Clustering

4.4.4 K-Medoids

Pada tahap clustering menggunakan k medoids harus melalui beberapa tahapan yang sama dengan k means seperti preprocessing data dan transformasi data. Berikut ini merupakan tahapan perancangan dan implementasi k-medoids :

1. Import library yang akan digunakan, seperti Pandas, Numpy dan Sklearn untuk menjalankan algoritma k means clustering.
2. Load data yang akan dianalisis
3. Melakukan tahapan preprocessing data.
4. Melakukan transformasi data
5. Menentukan jumlah cluster menggunakan metode elbow dan silhouette diagram
6. Inisialisasi model k means dengan jumlah cluster yang telah ditentukan
7. Visualisasi hasil k medoids.

Berikut ini merupakan Flowchart dari k medoids yang ditunjukan pada Gambar 4



Gambar 4 Flowchart K-Medoids

4.5 Evaluation

Setelah melakukan proses clustering menggunakan metode k means dan k medoids selanjutnya yaitu tahap *evaluation*, pada tahap *evaluation* menggunakan metode *DBI (davies bouldin index)* yang diperlukan untuk menilai kualitas clustering dan membandingkan antara k means dan k medoids. Untuk menjalankan metode *DBI* menggunakan fungsi *davies_bouldin_score()* adalah fungsi bawaan dari *library scikit-learn.*

4.6 Knowledge

Dalam tahap *knowledge* merupakan tahapan terakhir yaitu menarik informasi baru yang diperoleh dari hasil clustering dari metode k means maupun k medoids serta perbedaan dari kedua metode clustering tersebut.

BAB V
HASIL DAN PEMBAHASAN

5.1 Hasil

Pada hasil dari penelitian yang telah dilakukan terkait dengan perbandingan metode *k-means* dan *k-medoids* dalam melakukan clustering pada data Steam. Pada bab ini akan dijelaskan secara detail mengenai hasil eksperimen yang telah dilakukan, evaluasi hasil clustering menggunakan metode *Davies-Bouldin Index (DBI),* serta analisis dan interpretasi hasil *clustering* dari kedua metode tersebut. Hasil dan pembahasan yang disajikan diharapkan dapat memberikan gambaran yang jelas mengenai keefektifan dan keefisienan dari masing-masing metode clustering dalam melakukan pengelompokan data pada data Steam.

5.2 Deskripsi Data

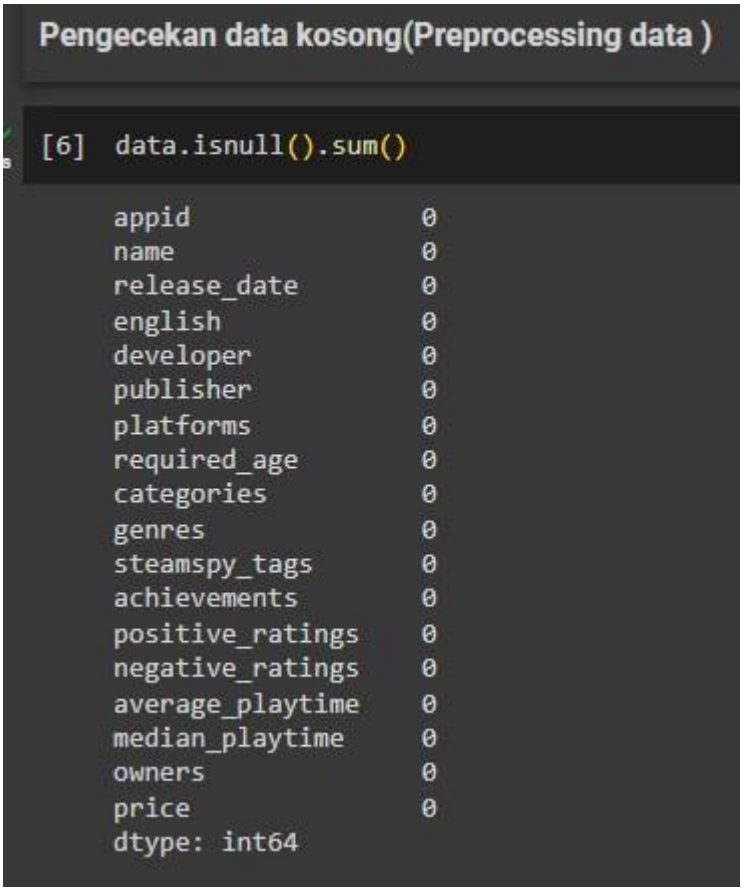
Data yang digunakan dalam penelitian ini merupakan data yang didapatkan dari kaggle dengan judul *Steam Store Games (Clean data)* data tersebut dikumpulkan sekitar Mei 2019, data ini berisikan informasi game yang ada pada steam dengan jumlah data 27075 data dan 18 kolom informasi, setelah melakukan observasi terdapat data kosong (missing value), oleh karena itu dilakukan tahap preprocessing. Untuk deskripsi tiap kolom dapat dilihat pada lampiran 1

5.3 Hasil Dari Preprocessing Data

Persiapan data merupakan langkah penting dalam analisis data yang dilakukan untuk membersihkan dan mengatur data sebelum dilakukan analisis lebih lanjut. Proses ini memiliki beberapa tahapan yang penting untuk memastikan hasil analisis yang akurat, berikut ini merupakan beberapa tahapan yang terjadi pada tahap preprocessing data.

5.3.1 Pengecekan Data Kosong

Tahap pertama pada preprocessing data yaitu pengecekan data kosong, pada tahap ini memeriksa apakah ada data kosong yang dapat mengganggu dalam menganalisa data. Dalam pengecekan data kosong tiap variabel menampilkan angka 0 yang menandakan tidak ditemukannya data kosong, menunjukan bahwa data memiliki representasi yang baik atau bisa diartikan juga dengan data steam memiliki informasi yang lengkap, untuk gambar dari hasil pengecekan data kosong dapat dilihat pada Gambar 5



Gambar 5 Pengecekan Data Kosong

5.3.2 Pengecekan Data Duplikat

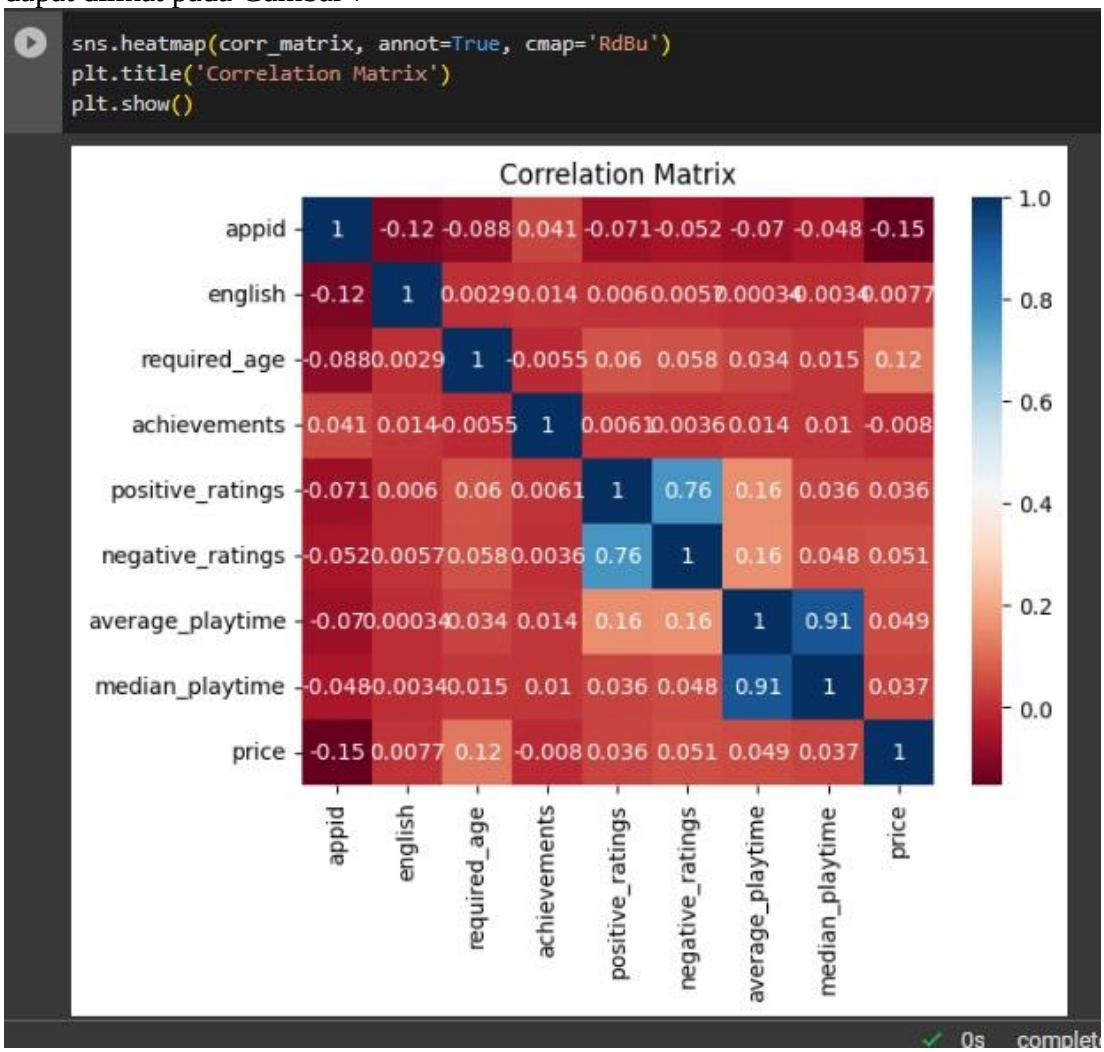
Setelah proses pengecekan data kosong maka selanjutnya yaitu pengecekan data duplikat, pada tahap ini memeriksa apakah ada data yang terindikasi memiliki duplikat yang dapat mengganggu dalam menganalisa data, dari hasil pengecekan tidak ditemukannya data duplikat, untuk pengecekan data duplikat bisa dilihat pada Gambar 6

Gambar 6 Pengecekan Data Duplikat

5.3.3 Pemilihan *Variabel*

pemilihan *variabel* menggunakan metode korelasi. Tujuan dari langkah ini adalah untuk mengukur hubungan antara dua variabel dan memilih variabel yang memiliki korelasi yang kuat. Dengan menggunakan metode korelasi, dapat menentukan apakah terdapat hubungan positif atau negatif antara dua *variabel* serta seberapa kuat hubungan tersebut. Hal ini membantu dalam pemilihan *variabel* yang paling relevan untuk proses clustering, dari hasil penggunaan korelasi menunjukan

hanya 6 *variabel* yang memiliki nilai rata rata hubungan yang baik dibandingkan dengan variabel lain. 6 variabel itu diantaranya *positive_ratings*, *negative_ratings*, *average_playtime*, *median_playtime*, *price dan english*, untuk pemilihan variabel dapat dilihat pada Gambar 7



Gambar 7 Hasil Metode Korelasi

5.3.4 Pengecekan *Outlier*

Tahap pengecekan data *outlier* menggunakan *boxplo*t Langkah ini bertujuan untuk mengidentifikasi dan menangani nilai-nilai yang ekstrem atau berbeda jauh dari pola umum data. Hasil pengecekan outlier pada 6 variabel menggunakan *boxplot* terindikasi memiliki outlier dengan persentase *outlier* yang beragam mulai dari 1.89% hingga 22.79%. kemunculan outlier dapat disebabkan oleh variasi alami dalam data atau perbedaan signifikan antara kelompok data. Jika rentang data sangat berbeda jauh, misalnya ada beberapa data dengan nilai yang sangat tinggi atau rendah, maka metode yang digunakan untuk mendeteksi outlier mungkin mengidentifikasi data tersebut sebagai outlier. Dalam penelitian ini, penghapusan outlier tidak diperlukan. Penyebab tidak dilakukan penghapusan outlier yaitu karena penulis meyakini tidak terbuktinya ada kesalahan pengukuran ataupun human error. Untuk hasil dari pengecekan outlier dapat dilihat pada Gambar 8.

```
Variabel positive_ratings: Total Outliers = 4286, Jumlah Data = 27075, Persentase Outliers = 15.83%
Variabel negative_ratings: Total Outliers = 3957, Jumlah Data = 27075, Persentase Outliers = 14.61%
Variabel average_playtime: Total Outliers = 6170, Jumlah Data = 27075, Persentase Outliers = 22.79%
Variabel median_playtime: Total Outliers = 6170, Jumlah Data = 27075, Persentase Outliers = 22.79%
Variabel price: Total Outliers = 1975, Jumlah Data = 27075, Persentase Outliers = 7.29%
Variabel english: Total Outliers = 511, Jumlah Data = 27075, Persentase Outliers = 1.89%
```

Gambar 8 Pengecekan Outlier

5.4 Hasil Dari *Data Transformation*

Pada tahap ini terjadi proses transformasi data numerik, salah satu teknik yang digunakan yaitu standar scaler dengan menggunakan fungsi "fit_transform()"
Setelah dilakukannya data transformation terjadi perubahan menjadi rentang nilai lebih kecil, hasil dari transformation dapat dilihat pada Gambar 9

Scaling Data (Transformation Data)

```
[70] scaler = StandardScaler()
     x_train=scaler.fit_transform(x_train)
     x_train

array([[ 0.13869593,  6.50574099,  0.73000595,  9.55782868,  0.07262355,
         0.14118582],
       [ 0.13869593,  0.12204528,  0.09847998,  0.06961946, -0.03571022,
        -0.2651749 ],
       [ 0.13869593,  0.12720634,  0.0436357 ,  0.02035849, -0.0476057 ,
        -0.2651749 ],
       ...,
       [ 0.13869593, -0.05269322, -0.04901611, -0.08199485, -0.0620502 ,
        -0.2651749 ],
       [ 0.13869593, -0.05258789, -0.04924949, -0.08199485, -0.0620502 ,
        -0.11278963],
       [ 0.13869593, -0.05248256, -0.04924949, -0.08199485, -0.0620502 ,
        -0.11278963]])
```

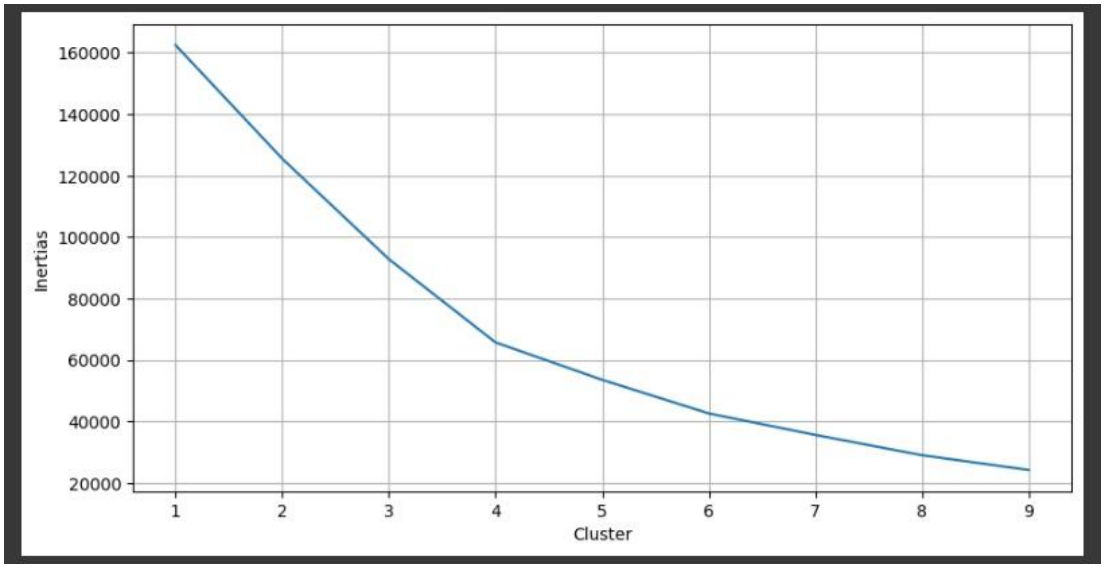
Gambar 9 Hasil dari Transformasi data

5.5 Hasil Dari *Data Mining*

Pada tahap data mining terjadi beberapa proses seperti penentuan jumlah cluster dengan metode *elbow* dan silhouette diagram setelah itu masuk ke tahap cluster menggunakan k means dan k medoids, dan berikut ini merupakan hasil dari proses data mining

5.5.1 Hasil Dari Metode *Elbow*

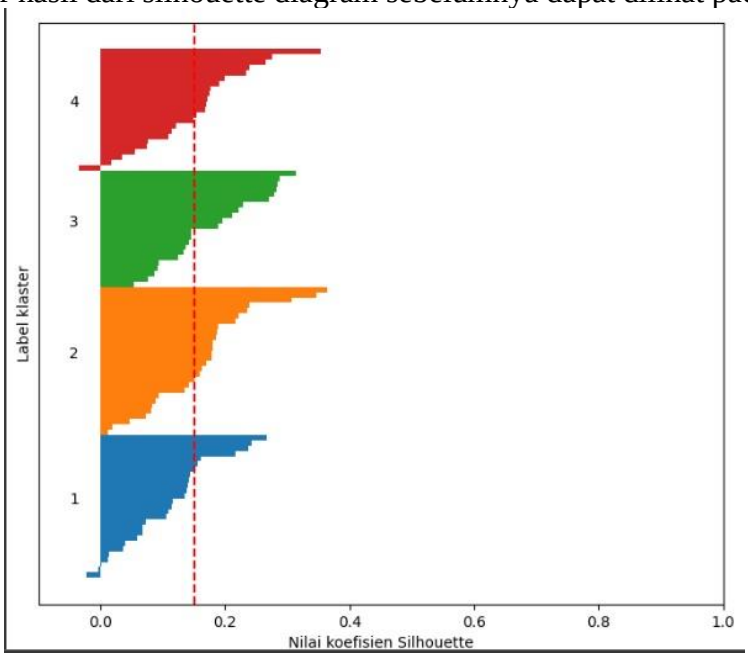
Metode *elbow* digunakan untuk menentukan jumlah *clustering* yang optimal Untuk mengetahui jumlah cluster yang optimal dengan cara membandingkan nilai dari tiap cluster, dan pada proses ini didapatkan nilai cluster optimal yaitu 4, dengan alasan cluster optimal yaitu memiliki nilai sum of error yang cukup kecil dan jumlah cluster yang kecil juga, di dalam penilaian melalui metode elbow cluster 4 memiliki nilai 65748 dengan posisi garis yang pas di lekukan sehingga cluster 4 merupakan cluster yang optimal, untuk gambar hasil dari metode elbow dapat dilihat pada Gambar 10.



Gambar 10 Metode Elbow

5.5.2 Hasil Silhouette Diagram

Untuk memastikan bahwa 4 cluster merupakan cluster optimal maka perlu pengecekan menggunakan *silhouette diagram* untuk menentukan jumlah cluster tersebut sudah optimal dapat dilihat melalui pola bar yang memiliki ukuran yang sama atau tidak terlalu jauh berbeda dan setiap bar terkena garis merah, apabila kriteria tersebut terpenuhi bisa dipastikan bahwa cluster 4 merupakan cluster yang optimal. Untuk gambar hasil dari silhouette diagram sebelumnya dapat dilihat pada Gambar 11



Gambar 11 Silhouette Diagram

5.5.3 Hasil *K-Means Clustering*

Pada hasil *clustering* menggunakan k means didapatkan hasil yaitu anggota cluster 1 berjumlah 26559 , anggota cluster 2 berjumlah 3, anggota cluster 3 berjumlah 2 dan cluster 4 berjumlah 511. Dan saat proses menjalankan algoritma k means ram yang digunakan yaitu 1,5 GB yang menandakan bahwa algoritma program cukup

ringan. Dari hasil *clustering* menggunakan k means terdapat informasi baru berupa pola atau ciri khusus dari anggota cluster 2, 3 dan 4. Berikut ini merupakan pola dari tiap cluster

Cluster 1 memiliki atribut dari variabel english, variabel ini berisikan angka 0 dan 1 yang artinya 0 adalah game yang tidak mendukung bahasa inggris, sedangkan 1 yaitu game yang mendukung bahasa inggris. Untuk cluster 1 memiliki ciri ciri yaitu semua anggotanya berisikan game yang mendukung bahasa inggris

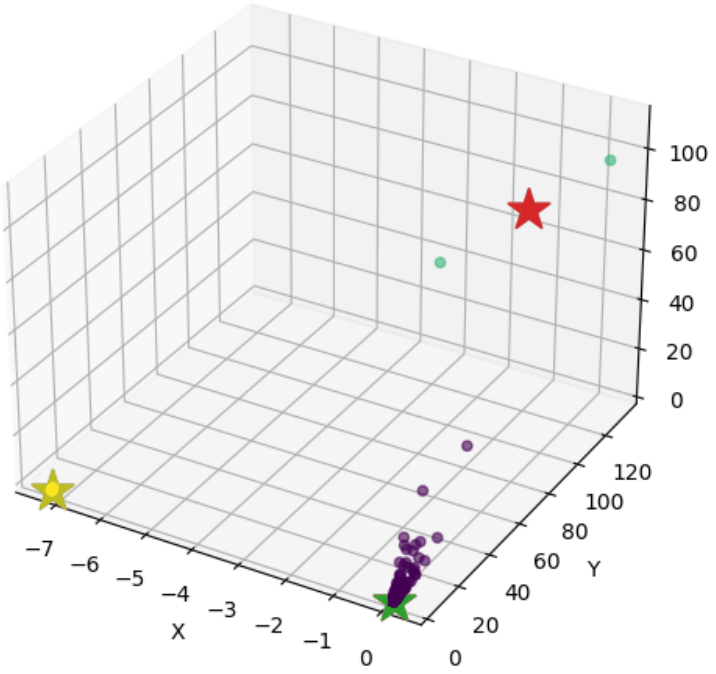
Cluster 2 memiliki atribut median_playtime, variabel ini menjelaskan informasi berupa rata-rata waktu pengguna permainan semua anggota dari cluster 2 memiliki nilai diatas 190000, cluster 2 mengindikasikan diisi oleh game yang paling sering dimainkan, dibandingkan dengan game lain anggota cluster 2 merupakan pemegang nilai tertinggi pada atribut median_playtime

Cluster 3 memiliki atribut positif_rating, variabel ini menjelaskan informasi berupa tanggapan positif dari para pemainnya, cluster 3 mengindikasikan diisi oleh game yang cukup populer karena anggota cluster 3 merupakan pemegang nilai rating positif paling tinggi dibandingkan dengan game lain. Selain itu cluster 3 memiliki ciri ciri lain seperti berkategori multiplayer, bergenre action serta dengan harga 0 yang menandakan bebas untuk diunduh.

Cluster 4 memiliki atribut english semua anggota cluster 4 memiliki nilai 0 pada variabel english yang menandakan bahwa game dari cluster 4 tidak mendukung bahasa inggris.

Tabel 2 Titik Pusat Cluster

Titik Pusat Cluster	Koordinat
Cluster 1	0.1387, -0.0054, -0.0070, -0.0087, -0.0099, 0.0012
Cluster 2	0.1387, -0.0363, -0.0279, 69.4511, 80.8780, -0.7718
Cluster 3	0.1387, 82.6449, 103.7335, 12.3515, 3.9603, -0.7718
Cluster 4	-7.2100, -0.0433, -0.0411, -0.0024, 0.0246, -0.0555



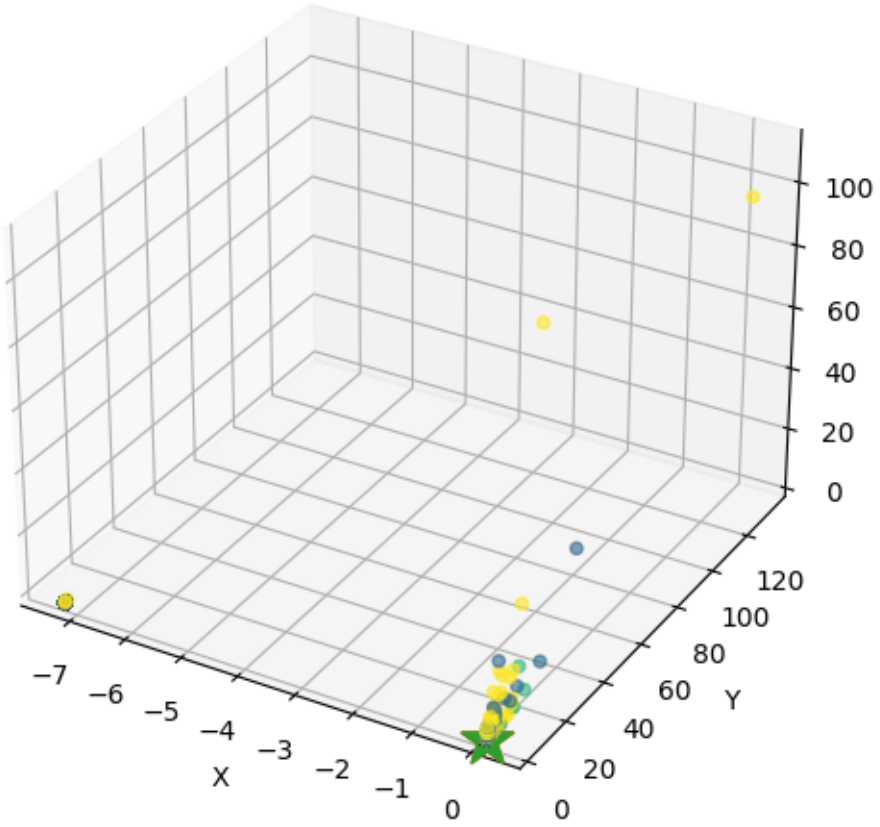
Gambar 12 Titik Pusat K-Means *Clustering*

5.5.4 Hasil K Medoids

Pada hasil *clustering* menggunakan k-medoids didapatkan hasil yaitu anggota cluster 0 yaitu berjumlah 10954, cluster 1 yaitu berjumlah 7149, cluster 2 berjumlah 9350, cluster 3 berjumlah 6680 dan cluster 4 3896. Saat proses menjalankan algoritma k medoids ram yang digunakan yaitu 8GB yang menandakan bahwa algoritma program cukup berat. Pola atau ciri khusus yang dihasilkan oleh k-medoids pada data steam yaitu memiliki pola pada rentang harga, cluster 1 semua anggotanya memiliki rentang harga mulai dari 4.49$ sampai 5.59$ sedangkan cluster 2 memiliki rentang harga 0 – 2.39$ untuk cluster 3 memiliki rentang harga 5.59 – 11.24$ yang terakhir cluster 4 memiliki rentang harga 9.99$ sampai 303.99$. diasumsikan cluster 2 diisi oleh game dengan harga murah sedangkan cluster 4 diisi oleh game dengan harga yang cukup mahal.

Tabel 3 Pusat Cluster K Medoids

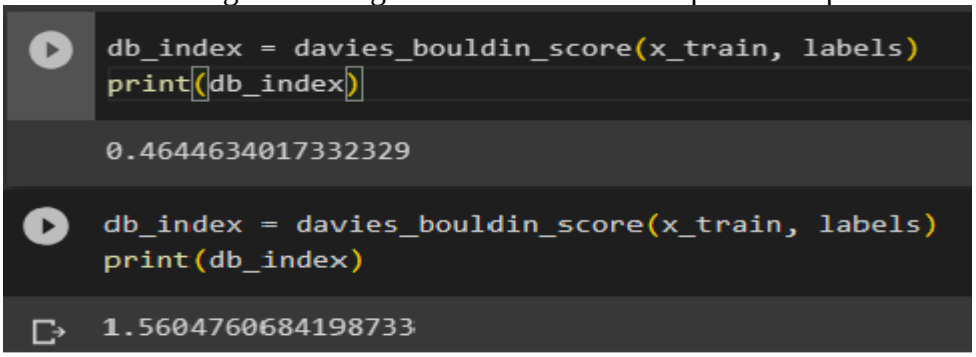
Titik Pusat Cluster	Koordinat
Cluster 1	0.1387, -0.0519, -0.0481, -0.0820, -0.0621, -0.2651
Cluster 2	0.1387, -0.0518, -0.0474, -0.0820, -0.0621, -0.6702
Cluster 3	0.1387, -0.0496, -0.0450, -0.0787, -0.0595, 0.1413
Cluster 4	0.1387, -0.0386, -0.0348, -0.0535, -0.0400, 1.1955



Gambar 13 Titik Pusat Cluster K Medoids

5.6 Pembahasan Hasil Evaluasi Menggunakan DBI

Berdasarkan hasil *DBI* yang diperoleh, nilai *DBI* untuk metode *k-means* yaitu 0.4644 lebih rendah dibandingkan dengan nilai *DBI* untuk metode k-medoids yaitu 1.5604 . Hal ini menunjukkan bahwa metode *k-means* memberikan hasil *clustering* yang lebih baik dibandingkan dengan metode *k-medoids* dalam hal kualitas *clustering*. Alasan k-means *clustering* lebih baik dari k-medoids karena memiliki nilai kurang dari satu, menandakan bahwa kualitas cluster cukup baik. Sedangkan untuk k-medoids memiliki hasil kurang bagus karena memiliki nilai diatas 1 yang menandakan kualitas cluster tidak terlalu bagus. Untuk gambar hasil dari *DBI* dapat dilihat pada Gambar 15



Gambar 14 Hasil Evaluasi Menggunakan DBI

BAB VI
Kesimpulan Dan saran

6.1 Kesimpulan

Setelah melakukan perbandingan antara metode *K-means* dan *K-medoids* pada data informasi game digital di platform Steam dapat diambil kesimpulan. Pada bab ini, akan diuraikan secara jelas dan singkat hasil analisis dan evaluasi yang telah dilakukan terhadap kedua metode *clustering* serta saran-saran yang dapat diberikan untuk penelitian selanjutnya. Kesimpulan yang diperoleh dari penelitian ini akan memberikan gambaran tentang kelebihan dan kekurangan masing-masing metode *clustering*.

Dari hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa data Steam dapat diolah menggunakan metode *clustering k-means* dan *k-medoids*. Pada tahap preprocessing data, terdapat beberapa tahapan yang meliputi data cleaning dan pengecekan data duplikat. Hasil pengecekan menunjukkan bahwa tidak ditemukan data kosong atau duplikat, yang menandakan bahwa data Steam memiliki representasi yang baik dan informasi yang lengkap. Selanjutnya, pada tahap pemilihan variabel menggunakan korelasi, terdapat 6 variabel yang dipilih karena memiliki nilai korelasi positif. Hal ini menunjukkan bahwa keenam variabel tersebut memiliki hubungan yang dapat membantu meningkatkan kualitas *clustering.*

Hasil pengecekan outlier pada 6 variabel menggunakan boxplot terindikasi memiliki outlier dengan persentase outlier yang beragam mulai dari 1.89% hingga 22.79%. kemunculan outlier dapat disebabkan oleh variasi alami dalam data atau perbedaan signifikan antara kelompok data. Jika rentang data sangat berbeda jauh, misalnya ada beberapa data dengan nilai yang sangat tinggi atau rendah, maka metode yang digunakan untuk mendeteksi outlier mungkin mengidentifikasi data tersebut sebagai outlier. Dalam penelitian ini, penghapusan outlier tidak diperlukan. Penyebab tidak dilakukan penghapusan outlier yaitu karena penulis meyakini tidak terbuktinya ada kesalahan pengukuran ataupun human error. Oleh karena itu penghapusan outlier tidak dilakukan.

Setiap *cluster* dari kedua metode memiliki pola atau ciri khusus pada metode *k-means* yaitu sebagai berikut :

1. Cluster 1 memiliki atribut dari *variabel english*, variabel ini berisikan angka 0 dan 1 yang artinya 0 adalah game yang tidak mendukung bahasa inggris, sedangkan 1 yaitu game yang mendukung bahasa inggris. Untuk cluster 1 memiliki ciri ciri yaitu semua anggotanya berisikan game yang mendukung bahasa inggris
2. Cluster 2 memiliki atribut *median_playtime*, variabel ini menjelaskan informasi berupa rata-rata waktu pengguna permainan semua anggota dari cluster 2 memiliki nilai diatas 190000, cluster 2 mengindikasikan diisi oleh game yang paling sering dimainkan, dibandingkan dengan game lain anggota cluster 2 merupakan pemegang nilai tertinggi pada atribut median_playtime
3. Cluster 3 memiliki atribut *positif_rating*, variabel ini menjelaskan informasi berupa tanggapan positif dari para pemainnya, cluster 3 mengindikasikan diisi oleh game yang cukup populer karena anggota cluster 3 merupakan pemegang nilai rating positif paling tinggi dibandingkan dengan game lain. Selain itu

cluster 3 memiliki ciri ciri lain seperti berkategori multiplayer, bergenre action serta dengan harga 0 yang menandakan bebas untuk diunduh.

4. Cluster 4 memiliki atribut *english* semua anggota cluster 4 memiliki nilai 0 pada variabel english yang menandakan bahwa game dari cluster 4 tidak mendukung bahasa inggris.

Sementara itu, pada metode K Medoids, setiap cluster memiliki ciri berdasarkan rentang harga. Cluster 1 memiliki rentang harga mulai dari 4,49$ hingga 5,59$, cluster 2 memiliki rentang harga 0 - 2,39$, cluster 3 memiliki rentang harga 5,59$ - 11,24$, dan cluster 4 memiliki rentang harga 9,99$ hingga 303,99$. Diasumsikan bahwa cluster 2 diisi oleh game dengan harga murah, sedangkan cluster 4 diisi oleh game dengan harga yang cukup mahal.

Dalam evaluasi kualitas *clustering,* digunakan *Davies-Bouldin Index (DBI)* untuk mengevaluasi kedua metode *clustering.* Nilai evaluasi menggunakan DBI menunjukkan bahwa k-means memiliki nilai 0,4644, sedangkan K Medoids memiliki nilai 1,5604. Semakin kecil nilai DBI yang diperoleh, menunjukan hasil cluster yang lebih baik. Berdasarkan nilai evaluasi tersebut, dapat disimpulkan bahwa k-means lebih baik daripada K Medoids dalam melakukan *clustering* pada kasus data Steam. Selain itu hasil pengelompokan menggunakan k-means lebih pariatif dibandingkan *k-medoids,* dari penggunaan ram k-means juga lebih unggul daripada K Medoids pada metode k-means penggunaan ram dibawah 2GB, sedangkan *k-medoids* memakan banyak ram yaitu 7 – 8GB penyebab tingginya penggunaan ram karena K Medoids lemah terhadap mengolah data dengan jumlah data yang banyak hal inilah yang mungkin menyebabkan tingginya penggunaan ram. Dengan mempertimbangkan semua faktor yang telah dibahas, dapat disimpulkan bahwa k-means *clustering* lebih unggul dalam mengolah data steam.

Selain itu, hasil pengelompokan menggunakan k-means lebih pariatif dibandingkan K Medoids. Dari penggunaan RAM, k-means juga lebih unggul daripada K Medoids pada metode k-means penggunaan RAM di bawah 2GB. Sedangkan K Medoids menggunakan banyak RAM yaitu 7-8GB. Penyebab tingginya penggunaan RAM karena K Medoids lemah terhadap mengolah data dengan jumlah data yang banyak. Hal ini mungkin menyebabkan tingginya penggunaan RAM. Dengan mempertimbangkan semua faktor yang telah dibahas, dapat disimpulkan bahwa k-means *clustering* lebih unggul dalam mengolah data steam.

6.2 Saran

Berdasarkan hasil analisis dan kesimpulan, dapat diberikan saran diantaranya sebagai berikut:

1. Dalam tahap preprocessing data, sebaiknya dilakukan pemilihan metode yang cocok untuk mendeteksi outlier seperti metode yang dapat menangani nilai ekstrim pada suatu data.
2. Sebaiknya menggunakan metode *clustering* yang lebih kompleks untuk meningkatkan interpretasi hasil *clustering.*

DAFTAR PUSTAKA

Ahuja, R., Solanki, A., & Anand, N. (2019). Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor. *Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor*.

Balmford, W. (2020). *Collecting, displaying and not playing: Steam sales and digital game collections*. https://www.pcgamer.com/analyst-says-digital-sales-made-up-92-percent-of-

Chunhui Yuan, H. Y. (2019). *Research on K-Value Selection Method of K-Means Clustering Algorithm*.

De Luisa, A., Hartman, J., Nabergoj, D., Pahor, S., Rus, M., Stevanoski, B., Demšar, J., & Štrumbelj, E. (2021). *Predicting the Popularity of Games on Steam*. http://arxiv.org/abs/2110.02896

Dewi, D. A. I. C., & Pramita, D. A. K. (2019). Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *Matrix : Jurnal Manajemen Teknologi Dan Informatika*, *9*(3), 102–109. https://doi.org/10.31940/matrix.v9i3.1662

Dhruv, A. J., Patel, R., & Doshi, N. (2021). *Python: The Most Advanced Programming Language for Computer Science Applications*. 292–299. https://doi.org/10.5220/0010307902920299

Febrian Sengkey, D., Diane Kambey, F., Paulus Lengkong, S., Reynaldo Joshua, S., & Valentino Florensius Kainde, H. (2020). Pemanfaatan Platform Pemrograman Daring dalam Pembelajaran Probabilitas dan Statistika di Masa Pandemi CoVID-19. *Jurnal Informatika*, *15*(4), 257–264.

Jumadi Dehotman Sitompul, B., Salim Sitompul, O., & Sihombing, P. (2019). Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm. *Journal of Physics: Conference Series*, *1235*(1). https://doi.org/10.1088/1742-6596/1235/1/012015

Kamila, I., & Khairunnisa, U. (2019). Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan. *Jurnal Ilmiah Rekayasa Dan Manajemen Sistem Informasi*, *5*(1), 119–125.

Karim, M. R., Beyan, O., Zappa, A., Costa, I. G., Rebholz-Schuhmann, D., Cochez, M., & Decker, S. (2021). Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, *22*(1), 393–415. https://doi.org/10.1093/bib/bbz170

Luthfi, E., Wahyu Wijayanto, A., & Statistika, P. (2021). *Analisis perbandingan metode hirearchical, k-means, dan k-medoids clustering dalam pengelompokkan indeks pembangunan manusia Indonesia*. *4*, 761–773. http://journal.feb.unmul.ac.id/index.php/INOVASI

Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile. *Entropy*, *23*(4). https://doi.org/10.3390/e23040485

Putra, R. R., & Wadisman, C. (2018). Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means. *INTECOMS: Journal of Information Technology and Computer Science*, *1*(1), 72–77. https://doi.org/10.31539/intecoms.v1i1.141

Putu, N., Merliana, E., & Santoso, A. J. (2018). *PROSIDING SEMINAR NASIONAL MULTI DISIPLIN ILMU & CALL FOR PAPERS UNISBANK (SENDI_U) Kajian Multi Disiplin Ilmu untuk Mewujudkan Poros Maritim dalam Pembangunan Ekonomi Berbasis Kesejahteraan Rakyat ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING*.

Rahman, F., Ridho, I. I., Muflih, M., Pratama, S., Raharjo, M. R., & Windarto, A. P. (2020). Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country. *IOP Conference Series: Materials Science and Engineering*, *835*(1). https://doi.org/10.1088/1757-899X/835/1/012058

Rao, W., Xia, J., Lyu, W., & Lu, Z. (2019). Interval data-based k-means clustering method for traffic state identification at urban intersections. *IET Intelligent Transport Systems*, *13*(7), 1106–1115. https://doi.org/10.1049/iet-its.2018.5379

Soni Madhulatha, T. (2012). *AN OVERVIEW ON CLUSTERING METHODS*. *2*(4), 719–725. www.iosrjen.org

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *4*(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Velu, A., & Whig, P. (2021). *Impact of Covid Vaccination on the Globe using data analytics A R T I C L E I N F O ABSTRACT*. www.ijsdcs.com

Xia, S., Peng, D., Meng, D., Zhang, C., Wang, G., Giem, E., Wei, W., & Chen, Z. (2022). Ball k-Means: Fast Adaptive Clustering With No Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(1), 87–99. https://doi.org/10.1109/TPAMI.2020.3008694

LAMPIRAN

Lampiran 1 Deskripsi Variabel Data Steam

Appid	: Pengidentifikasi unik untuk setiap judul
Name	: Nama/judul (game)
release_date	: Tanggal rilis game
english	: Dukungan bahasa: 1 jika dalam bahasa Inggris
developer	: Nama pengembang
publisher	: Nama Penerbit
platforms	: Daftar platform yang didukung
required_age	: Usia minimum yang disyaratkan menurut standar PEGI UK
categories	: Daftar kategori game pemain tunggal; multipemain
genres	: Daftar genre game
steamspy_tags	: Daftar tag game steamspy,mirip dengan genre tetapi dipilih oleh komunitas.
Achievements	: Jumlah pencapaian dalam game
positive_ratings	: Jumlah peringkat positif, dari SteamSpy
negative_ratings	: Jumlah peringkat negatif, dari SteamSpy
average_playtime	: Rata-rata waktu bermain pengguna, dari SteamSpy
median_playtime	: Waktu bermain pengguna rata-rata, dari SteamSpy
owners	: Perkiraan jumlah pemilik game
price	: Harga penuh judul saat ini dalam GBP, (pound sterling)

appid	name	release_date	english	developer	publisher	platforms	required_age
10	Counter-S	November 1, 2000	1	Valve	Valve	windows;mac;linux	0
20	Team Fort	April 1, 1999	1	Valve	Valve	windows;mac;linux	0
30	Day of De	May 1, 2003	1	Valve	Valve	windows;mac;linux	0
40	Deathmat	June 1, 2001	1	Valve	Valve	windows;mac;linux	0
50	Half-Life:	November 1, 1999	1	Gearbox Software	Valve	windows;mac;linux	0
60	Ricochet	November 1, 2000	1	Valve	Valve	windows;mac;linux	0
70	Half-Life	November 8, 1998	1	Valve	Valve	windows;mac;linux	0
80	Counter-S	March 1, 2004	1	Valve	Valve	windows;mac;linux	0
130	Half-Life:	June 1, 2001	1	Gearbox Software	Valve	windows;mac;linux	0
220	Half-Life 2	November 16, 2004	1	Valve	Valve	windows;mac;linux	0
240	Counter-S	November 1, 2004	1	Valve	Valve	windows;mac;linux	0
280	Half-Life:	June 1, 2004	1	Valve	Valve	windows;mac;linux	0
300	Day of De	July 12, 2010	1	Valve	Valve	windows;mac;linux	0
320	Half-Life 2	November 1, 2004	1	Valve	Valve	windows;mac;linux	0
340	Half-Life 2	October 27, 2005	1	Valve	Valve	windows;mac;linux	0

categories	genres	steamspy_tags	achievements	positive_ratings	negative_ratings	average_playtime	median_playtime	owners	price
Multi-player;(	Action	Action;FPS;Multiplay	0	124534	3339	17612	317	20000000	7,19
Multi-player;(	Action	Action;FPS;Multiplay	0	3318	633	277	62	10000000	3,99
Multi-player;'	Action	FPS;World War II;Mu	0	3416	398	187	34	10000000	3,99
Multi-player;(	Action	Action;FPS;Multiplay	0	1273	267	258	184	10000000	3,99
Single-player	Action	FPS;Action;Sci-fi	0	5250	288	624	415	10000000	3,99
Multi-player;(	Action	Action;FPS;Multiplay	0	2758	684	175	10	10000000	3,99
Single-player	Action	FPS;Classic;Action	0	27755	1100	1300	83	10000000	7,19
Single-player	Action	Action;FPS;Multiplay	0	12120	1439	427	43	20000000	7,19
Single-player	Action	FPS;Action;Sci-fi	0	3822	420	361	205	10000000	3,99
Single-player	Action	FPS;Action;Sci-fi	33	67902	2419	691	402	20000000	7,19
Multi-player;(	Action	Action;FPS;Multiplay	147	76640	3497	6842	400	20000000	7,19
Single-player	Action	FPS;Action;Sci-fi	0	3767	1053	190	214	5000000	0
Multi-player;(	Action	FPS;World War II;Mu	54	10489	1210	1356	134	10000000	7,19
Multi-player;'	Action	Action;FPS;Multiplay	0	6020	787	311	32	20000000	3,99
Single-player	Action	FPS;Action;Singlepla	0	5783	1020	46	29	20000000	0

Lampiran 2. Library Python yang digunakan

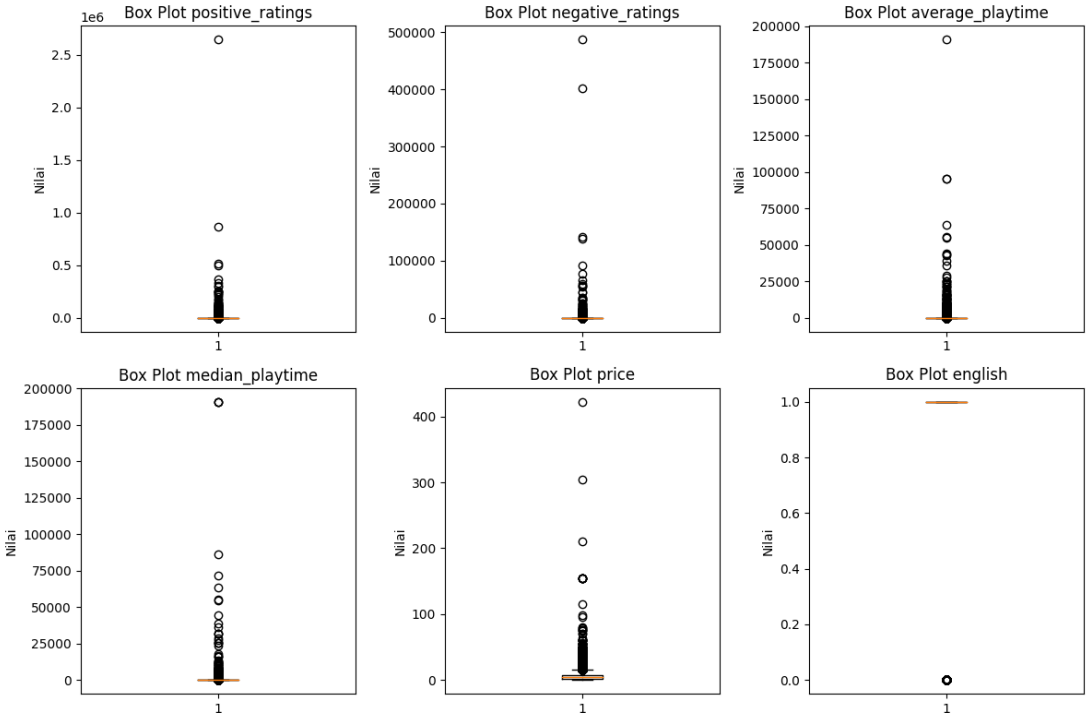
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import sklearn.cluster as cluster
from sklearn.cluster import KMeans
from sklearn_extra.cluster import KMedoids
from sklearn.metrics import davies_bouldin_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_samples, silhouette_score
```

Lampiran 3. Tipe Data Tiap Variabel

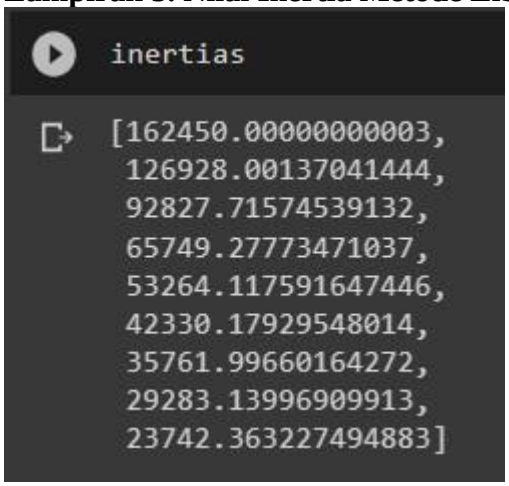
```
steam.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27075 entries, 0 to 27074
Data columns (total 18 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   appid             27075 non-null  int64  
 1   name              27075 non-null  object 
 2   release_date      27075 non-null  object 
 3   english           27075 non-null  int64  
 4   developer         27075 non-null  object 
 5   publisher         27075 non-null  object 
 6   platforms         27075 non-null  object 
 7   required_age      27075 non-null  int64  
 8   categories        27075 non-null  object 
 9   genres            27075 non-null  object 
 10  steamspy_tags     27075 non-null  object 
 11  achievements      27075 non-null  int64  
 12  positive_ratings  27075 non-null  int64  
 13  negative_ratings  27075 non-null  int64  
 14  average_playtime  27075 non-null  int64  
 15  median_playtime   27075 non-null  int64  
 16  owners            27075 non-null  object 
 17  price             27075 non-null  float64
dtypes: float64(1), int64(8), object(9)
memory usage: 3.7+ MB
```

Lampiran 4. Pengecekan Outlier Dengan Boxplot



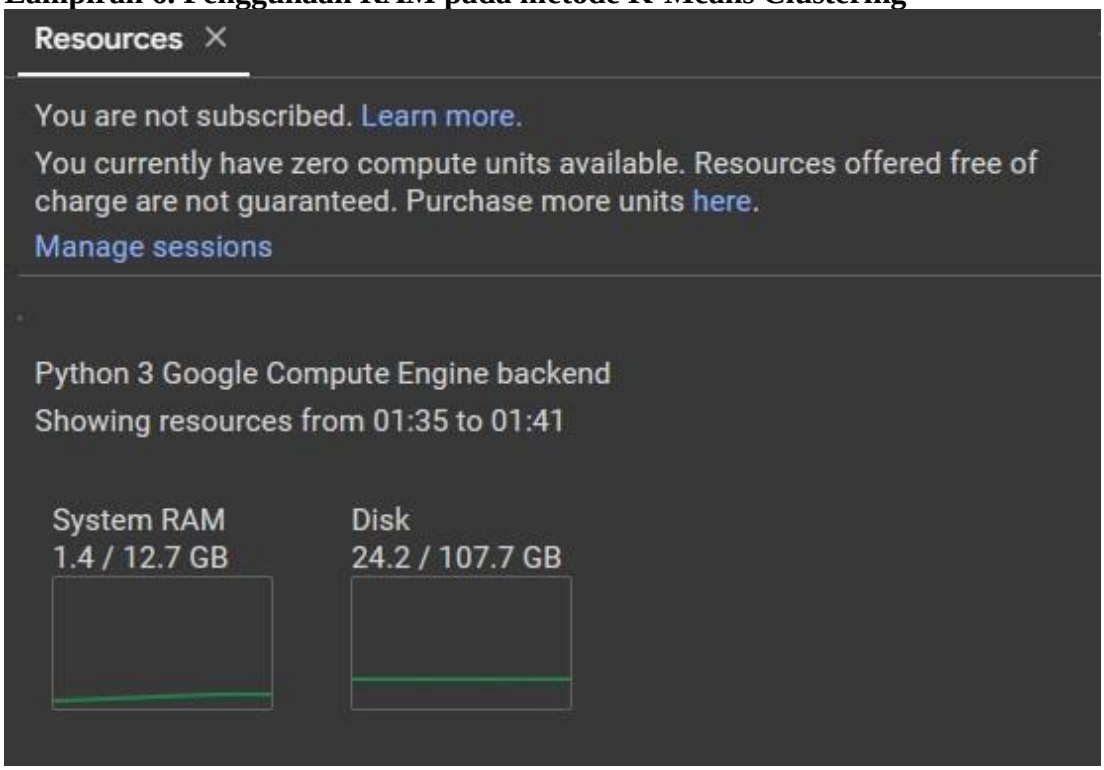
Lampiran 5. Nilai Inertia Metode Elbow



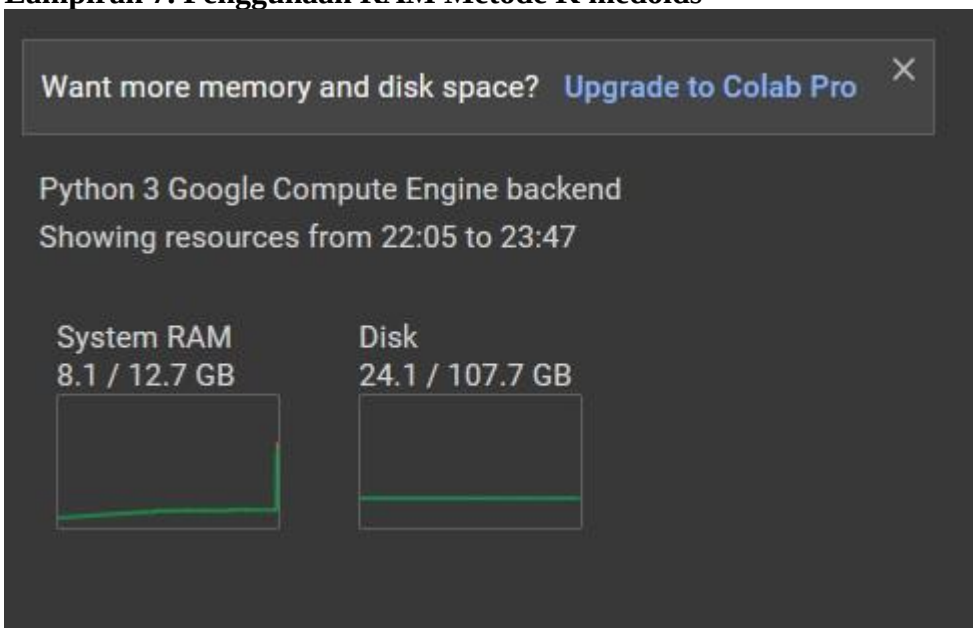
```
inertias

[162450.00000000003,
 126928.00137041444,
 92827.71574539132,
 65749.27773471037,
 53264.117591647446,
 42330.17929548014,
 35761.99660164272,
 29283.13996909913,
 23742.363227494883]
```

Lampiran 6. Penggunaan RAM pada metode K-Means Clustering



Lampiran 7. Penggunaan RAM Metode K medoids



Lampiran 8 Perhitungan manual K-means

Menghitung Centroid

Langkah pertama yaitu mengambil 3 data secara acak untuk dijadikan centroid awal setelah itu menggunakan rumus Euclidean Distence untuk mendapatkan jarak minimum data terhadap centroid.

Centroid awal

Centroid						
C1	3318	633	277	62	5000000	3,99
C2	1273	267	258	184	5000000	3,99
C3	12120	1439	427	43	10000000	7,19

Berikut ini adalah rumus Euclidean Distance

Berikut rumus Euclidean Distance :	$[(x, y), (a, b)] = \sqrt{(x-a)^2 + (y-b)^2}$

Menghitung jarak antar centroid

Iterasi pertama jarak data centroid 1

$\sqrt{(124534 - 3318)^2}$ + (3339– 633)^2 + (17612 – 277)^2 + (317– 62)^2 + (10000000 – 5000000)^2 + (7.19 – 3.99)^2 = 5001500

$\sqrt{(3318 - 3318)^2}$ + (633– 633)^2 + (277– 277)^2 + (62– 62)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 0

$\sqrt{(3416 - 3318)^2}$ + (398– 633)^2 + (187– 277)^2 + (34– 62)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 271,5014

$\sqrt{(1273 - 3318)^2}$ + (267– 633)^2 + (258– 277)^2 + (184– 62)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 2081,159773

$\sqrt{(5250 - 3318)^2}$ + (288– 633)^2 + (624– 277)^2 + (415– 62)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 2024,02248

(seterusnya sampai n data)

Iterasi pertama jarak data dengan centroid 2

$\sqrt{(124534 - 1273)^2}$ + (3339– 267)^2 + (17612 – 258)^2 + (317– 184)^2 + (10000000 – 5000000)^2 + (7.19 – 3.99)^2 = 5001550

$\sqrt{(3318 - 1273)^2}$ + (633– 267)^2 + (277– 258)^2 + (62– 184)^2 + (5000000 – 5000000)^2 + (7.19 – 3.99)^2 = 2081,16

$\sqrt{(3416 - 1273)^2}$ + (398– 267)^2 + (187– 258)^2 + (34– 184)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 2153,405

$\sqrt{(1273 - 1273)^2}$ + (267– 267)^2 + (358– 258)^2 + (184– 184)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 0

$\sqrt{(5250 - 1273)^2}$ + (288– 267)^2 + (624– 258)^2 + (415– 184)^2 + (5000000 – 5000000)^2 + (3.99 – 3.99)^2 = 4000,535839

(seterusnya sampai n data)

Iterasi pertama jarak data dengan centroid 3

$\sqrt{(124534 - 12120)^2}$ + (3339– 1439)^2 + (17612 – 427)^2 + (317– 43)^2 + (10000000 – 10000000)^2 + (7.19 – 719)^2 = 113736,1715

$\sqrt{(3318 - 12120)^2}$ + (633– 1439)^2 + (277– 427)^2 + (62– 43)^2 + (5000000 – 10000000)^2 + (3.99 – 719)^2 = 5000007,815

$\sqrt{(3416 - 12120)^2}$ + (398– 1439)^2 + (187– 427)^2 + (34– 43)^2 + (5000000 – 10000000)^2 + (3.99 – 719)^2 = 5000007,69

$\sqrt{(1273 - 12120)^2}$ + (267– 1439)^2 + (258– 427)^2 + (184– 43)^2 + (5000000 – 10000000)^2 + (3.99 – 719)^2 = 5000011,908

$\sqrt{(5250 - 12120)^2}$ + (288– 1439)^2 + (624– 427)^2 + (415– 43)^2 + (5000000 – 10000000)^2 + (3.99 – 719)^2 = 5000004,87

(seterusnya sampai n data)

Lampiran 9 Perhitungan manual k-medoids

Menghitung Centroid

Sama dengan sebelumnya pada K-means yaitu mengambil 2 data secara acak untuk dijadikan centroid awal setelah itu menggunakan rumus Euclidean Distence untuk mendapatkan jarak minimum data terhadap centroid.

Centroid awal K-medoids

Centroid						
C1	5250	288	624	415	5000000	3.99
C2	620	787	311	32	10000000	3.99

Menghitung Jarak Antar Centroid

Iterasi pertama jarak antar C1

Iterasi pertama jarak data centroid 1

√(124534-3318)2 + (3339– 633)2 + (17612 – 277)2 + (317– 62)2 + (10000000 – 5000000)2 + (7.19 – 3.99)2 = 5001500

√(3318-3318)2 + (633– 633)2 + (277– 277)2 + (62– 62)2 + (5000000 – 5000000)2 + (3.99 – 3.99)2 = 0

√(3416-3318)2 + (398– 633)2 + (187– 277)2 + (34– 62)2 + (5000000 – 5000000)2 + (3.99 – 3.99)2 = 271,5014

√(1273-3318)2 + (267– 633)2 + (258– 277)2 + (184– 62)2 + (5000000 – 5000000)2 + (3.99 – 3.99)2 = 2081,159773

√(5250-3318)2 + (288– 633)2 + (624– 277)2 + (415– 62)2 + (5000000 – 5000000)2 + (3.99 – 3.99)2 = 2024,02248

(seterusnya sampai n data)

Lampiran 10 Kartu bimbingan

Kartu Bimbingan Mahasiswa
Program Studi Ilmu Komputer FMIPA - UNPAK

Nama Mahasiswa : Candra Wisantra
NPM : 065116044
Judul Skripsi : Implementasi Metode K-means Clustering Dan K-Medoids Pada Data Steam
Pembimbing I : Dr. Prihastuti Harsani, M.Si.
Pembimbing II : Dr.Fajar Deli W.,SSI.MM.,M.Kom

No.	Hari, tanggal	Catatan	Tanda Tangan	
			Pemb. I	Pemb. II
1.	Selasa, 15 November 2022	Pengajuan judul pembimbing utama dan pendamping	1	2
2.	Kamis, 17 November 2022	Pemilihan variabel yang akan digunakan pada dataset Steam		
3.	Sabtu, 19 November 2022	Penambahan metode dalam pembersihan dataset	3	4
4.	Senin, 21 November 2022	Bimbingan proposal Refrensi diambil kurang dari 5 tahun kebelakang		
5.	Rabu, 23 November 2022	Bimbingan Proposal Pada BAB II mencantumkan link sumber dataset	5	6
6.	senin, 28 November 2022	Memindahkan perhitungan manual clustering dari BAB IV ke lampiran		
7.	Rabu, 15 Maret 2023	Pemeriksaan kembali mengenai kata asing yang digunkan supaya dicetak miring	7	8
8.	Jumat, 17 Maret 2023	Pemeriksaan kembali mengenai ukuran tabel dan grafik supaya lebih jelas		
9.	Senin, 20 Maret 2023	Bimbingan hasil BAB V menambahkan keterangan dari grafik clustering	9	10
10.	Jumat, 19 May 2023	Penambahan deskripsi mengenai dataset yang digunkan		
11.	Selasa, 13 Juni 2023	Penambahan Metode dalam menentukan jumlah cluster terbaik	11	12
12.	Rabu, 21 Juni 2023	Pada BAB V menambahkan tabel dan grafik titik pusat cluster		
13.	Senin, 3 juli 2023	Pada BAB VI menambahkan kembali kesimpulan mengenai hasil dari cluster	13	14
14.	Rabu, 5 Juli 2023	Pemeriksaan kembali dalam pemilhan kata yang digunakan dengan menggunakan kata baku yg baik		
15.			15	16
16.				
17.			17	18
18.				

Bogor, 13 Juli 2023
Ketua Program Studi Ilmu Komputer
FMIPA-UNPAK

Arie Qur'ania, M.Kom

Lampiran 11 Surat keputusan

YAYASAN PAKUAN SILIWANGI
Universitas Pakuan
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
Unggul, Mandiri & Berkarakter Dalam Bidang MIPA

KEPUTUSAN DEKAN
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PAKUAN No.: 4376/D/FMIPA/XII/2022

T E N T A N G

PENGANGKATAN PEMBIMBING TUGAS AKHIR
PADA PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PAKUAN

DEKAN FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PAKUAN,

Menimbang : a. bahwa setiap mahasiswa tingkat akhir Program Strata Satu (S1) harus melaksanakan Tugas Akhir sebagaimana tercantum di dalam kurikulum setiap Program Studi di lingkungan Fakultas MIPA Universitas Pakuan.
b. bahwa untuk pelaksanaan Tugas Akhir diperlukan pengawasan dari pembimbing.
c. bahwa sehubungan dengan point a dan b di atas perlu dituangkan dalam suatu Keputusan Dekan.

Mengingat : 1. Undang-undang RI No.: 20 Tahun 2003 tentang Sistem Pendidikan Nasional.
2. Peraturan Pemerintah No.: 60 Tahun 1999 tentang Pendidikan Tinggi.
3. Statuta Universitas Pakuan Tahun 2019.
4. Surat Keputusan Rektor Nomor: 35/KEP/REK/VIII/2020 tanggal 03 Agustus 2020 tentang Pemberhentian Dekan dan Wakil Dekan Masa Bakti 2015-2020 serta Pengangkatan Dekan dan Wakil Dekan Masa Bakti 2020-2025 di lingkungan Universitas Pakuan.
5. Ketentuan Akademik yang tercantum dalam Buku Panduan Studi Fakultas MIPA, Universitas Pakuan Tahun 2021.

Memperhatikan : Usulan dari Ketua Program Studi Ilmu Komputer FMIPA UNPAK.

M E M U T U S K A N

Menetapkan :

Pertama : Mengangkat pembimbing yang namanya tersebut di bawah ini :
1. Pembimbing Utama : Dr. Prihastuti Harsani, M.Si.
2. Pembimbing Pendamping : Fajar Delli W.,SSi.,MM.,M.Kom

Untuk membimbing dalam rangka melaksanakan tugas akhir bagi mahasiswa :
Nama : Candra Wisantra
NPM : 065116044
Program Studi : Ilmu Komputer
Judul Skripsi : Implementasi Metode K-Means Clustering dan K-Medoids Pada Dataset Steam

Jl. Pakuan P.O. Box 452, Bogor 16143, Telp./Fax. (0251) 8375547
Website : https://fmipa.unpak.ac.id

Kedua : Kepada para pembimbing diharapkan dapat menjalankan tugasnya sebagai pembimbing dengan sebaik-baiknya.

Ketiga : Dalam waktu 1 (satu) bulan setelah diterbitkannya SK ini, mahasiswa wajib melaksanakan Seminar Rencana Penelitian yang diselenggarakan oleh Program Studi Ilmu Komputer dengan dihadiri oleh Pembimbing dan Penguji.

Keempat : Dana untuk honorarium pembimbing dibebankan kepada mahasiswa yang ketentuannya diatur oleh Fakultas MIPA.

: Surat Keputusan ini berlaku untuk jangka waktu 1 (satu) tahun sejak tanggal ditetapkan sampai dengan mahasiswa tersebut Lulus Sidang/Ujian Skripsi, dengan ketentuan akan diadakan perubahan/perbaikan sebagaimana mestinya bila dikemudian hari terdapat kekeliruan dalam penetapannya.

Ditetapkan di : Bogor
Pada tanggal : 05 Desember 2022

Dekan,

Asep Denih, S,Kom., M.Sc., Ph.D.

Tembusan :

Kelima