PERBANDINGAN METODE MANHATTAN DAN EUCLIDEAN DALAM ANALISIS DATA RASIO GINI SUMATERA UTARA MENGGUNAKAN ALGORITMA K-MEANS

SKRIPSI

DISUSUN OLEH

MHD HASAN PASARIBU

NPM. 2009010113

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2024

PERBANDINGAN METODE MANHATTAN DAN EUCLIDEAN DALAM ANALISIS DATA RASIO GINI SUMATERA UTARA MENGGUNAKAN ALGORITMA K-MEANS

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer (S.Kom) dalam Program Studi Sistem Informasi pada Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Muhammadiyah Sumatera Utara

MHD HASAN PASARIBU

NPM. 2009010113

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI

UNIVERSITAS MUHAMMADIYAH SUMATERA UTARA

MEDAN

2024

LEMBAR PENGESAHAN

Judul Skripsi	:	PERBANDINGAN METODE MANHATTAN DAN EUCLIDEAN DALAM ANALISIS DATA RASIO GINI SUMATERA UTARA MENGGUNAKAN ALGORITMA K-MEANS
Nama Mahasiswa	:	MHD HASAN PASARIBU
NPM	:	2009010113
Program Studi	:	SISTEM INFORMASI

Menyetujui

Komisi Pembimbing

(Martiano, S.Pd., S.Kom., M.Kom)

NIDN. 0128029302

Ketua Program Studi

(Martiano, S.Pd., S.Kom., M.Kom)
NIDN. 0128029302

Dekan

(Dr. Al-Khowarizmi, S.Kom., M.Kom.)
NIDN. 0127099201

PERNYATAAN ORISINALITAS

PERBANDINGAN METODE MANHATTAN DAN EUCLIDEAN DALAM ANALISIS DATA RASIO GINI SUMATERA UTARA MENGGUNAKAN ALGORITMA K-MEANS

SKRIPSI

Saya menyatakan bahwa karya tulis ini adalah hasil karya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing disebutkan sumbernya.

Medan, 07 Mei 2024

Yang membuat pernyataan

Mhd Hasan Pasaribu

NPM. 2009010113

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika Universitas Muhammadiyah Sumatera Utara, saya bertanda tangan dibawah ini:

Nama	: Mhd Hasan Pasaribu
NPM	: 2009010113
Program Studi	: Sistem Informasi
Karya Ilmiah	: Skripsi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Muhammadiyah Sumatera Utara Hak Bebas Royalti Non-Eksekutif *(Non-Exclusive Royalty free Right)* atas penelitian skripsi saya yang berjudul:

PERBANDINGAN METODE MANHATTAN DAN EUCLIDEAN DALAM ANALISIS DATA RASIO GINI SUMATERA UTARA MENGGUNAKAN ALGORITMA K-MEANS

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksekutif ini, Universitas Muhammadiyah Sumatera Utara berhak menyimpan, mengalih media, memformat, mengelola dalam bentuk database, merawat dan mempublikasikan Skripsi saya ini tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis dan sebagai pemegang dan atau sebagai pemilik hak cipta.

Demikian pernyataan ini dibuat dengan sebenarnya.

Medan, 07 Mei 2024
Yang membuat pernyataan

Mhd Hasan Pasaribu
NPM. 2009010113

RIWAYAT HIDUP

DATA PRIBADI

Nama Lengkap	: Mhd Hasan Pasaribu
Tempat dan Tanggal Lahir	: Bandung 15 Mei 2001
Alamat Rumah	: Jl. Jermal IV no. 51, Medan Denai
Telepon/Faks/HP	: 0813-6052-0134
E-mail	: mhdhasanpasaribu@gmail.com
Instansi Tempat Kerja	: - (Belum Bekerja)
Alamat Kantor	: - (Belum Bekerja)

DATA PENDIDIKAN

SD	: SD SWASTA MUHAMMADIYAH 02	TAMAT: 2013
SMP	: SMP SWASTA PERTIWI MEDAN	TAMAT: 2016
SMA	: SMA NEGERI 03 MEDAN	TAMAT: 2019

KATA PENGANTAR

Rasa Syukur penulis panjatkan atas kehadirat Allah SWT karena berkah, rahmat dan karunian-nya saya dapat menyelesaikan skripsi ini yang merupakan salah satu persyaratan akademik untuk menyelesaikan pendidikan S1 jurusan Sistem Informasi. Adapun judul peneliti ialah: "Perbandingan Metode Manhattan Dan Euclidean Dalam Analisis Data Rasio Gini Sumatera Utara Menggunakan Algoritma K-Means".

Penghargaan dan terimakasih yang setulus-tulusnya kepada kedua orangtua saya yang sangat saya sayangi yang telah mencurahkan segenap cinta dan kasih sayang serta perhatian baik secara moral maupun materil. Semoga Allah SWT selalu melimpahkan seluruh berkahnya di dunia dan di akhirat atas budi baik yang telah diberikan kepada saya. Penulis tentunya berterima kasih kepada berbagai pihak dalam dukungan serta doa dalam penyelesaian skripsi. Penulis juga mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Agussani, M.AP., Rektor Universitas Muhammadiyah Sumatera Utara (UMSU).
2. Bapak Dr. Al-Khowarizmi, S.Kom., M.Kom. Dekan Fakultas Ilmu Komputer dan Teknologi Informasi (FIKTI) UMSU.
3. Bapak Martiano, S.Pd., S.Kom., M.Kom. Ketua Program Studi Sistem Informasi.
4. Ibu Yoshida Sary, S.Kom., M.Kom. Sekretaris Program Studi Sistem Informasi.
5. Bapak Martiano, S.Pd., S.Kom., M.Kom. Dosen Pembimbing Penulis Skripsi.
6. Bapak Prof. Dr. Fajar Pasaribu, S.E., M.Si. Selaku Ayah dari Penulis.
7. Ibu Prof. Dr. Widia Astuty, SE., M.Si., Ak., CA., CPAi. Selaku Ibu dari Penulis.
8. Semua pihak yang terlibat langsung ataupun tidak langsung yang tidak dapat penulis ucapkan satu-persatu yang telah membantu penyelesaian skripsi ini.

PERBANDINGAN METODE MANHATTAN DAN EUCLIDEAN DALAM ANALISIS DATA RASIO GINI SUMATERA UTARA MENGGUNAKAN ALGORITMA K-MEANS

ABSTRAK

Metode jarak Manhattan dan Euclidean adalah dua metode jarak yang memiliki kemampuan yang luar biasa dan cukup populer dikalangan peneliti, kedua metode ini cendrung digunakan oleh peneliti untuk memperoleh hasil pengklasteran dari data yang diolah dengan tujuan agar mengetahui pola tertentu dan dapat menghasilkan penafsiran-penafsiran tertentu, namun dari dua metode tersebut perlu dicari tahu metode yang mana yang paling baik performanya. Pada penelitian ini, peneliti akan melakukan pengujian metode jarak Manhattan dan Euclidean dengan menggunakan algoritma *K-means Clustering* terhadap data rasio Gini di Sumatera Utara mulai dari tahun 2000-2023 dengan menggunakan *library Pandas, Matplotlib,* dan *Seaborn*. Dalam proses penelitian ini, peneliti melakukan perhitungan hasil nilai performa yang terdiri dari *Dunn Index* dan jumlah iterasi pada dataset rasio gini pada setiap metode jarak. Berdasarkan hasil pengujian, didapatkan performa pada metode *K-Means Clusteri*ng yang menggunakan *Manhattan Distance* bernilai *Dunn Index* sebesar 0.2860941592407533 dengan jumlah iterasi sebanyak 3 iterasi, sedangkan pada metode *K-Means Clusteri*ng yang menggunakan *Euclidean Distance* bernilai *Dunn Index* sebesar 0.28572673870340515 dengan jumlah iterasi sebanyak 3 iterasi. Performa metode jarak dalam proses K-means Clustering terbaik pada dataset rasio Gini yaitu metode *Manhattan Distance*.

Kata Kunci: K-means; Euclidean Distance; Manhattan Distance; Rasio Gini.

COMPARISON OF MANHATTAN AND EUCLIDEAN METHODS IN THE ANALYSIS OF NORTH SUMATERA'S GINI RATIO DATA USING THE K-MEANS ALGORITHM

ABSTRACT

The Manhattan and Euclidean distance methods are two distance methods that have excellent abilities and are popular among researchers, these two methods tend to be used by researchers to obtain clustering results from processed data with the aim of knowing certain patterns and can produce certain interpretations, but from the two methods it is necessary to find out which method performs best. In this study, the researcher will test the Manhattan and Euclidean distance methods using the K-means Clustering algorithm on Gini ratio data in North Sumatera starting from 2000-2023 using the Pandas, Matplotlib, and Seaborn libraries. In this research process, the researcher calculated the results of the performance value consisting of the Dunn Index and the number of iterations in the Gini ratio dataset in each distance method. Based on the test results, the performance of the K-Means Clustering method using Manhattan Distance is resulting the Dunn Index value of 0.2860941592407533 with a total of 3 iterations, while the K-Means Clustering method using Euclidean Distance is resulting Dunn Index value of 0.28572673870340515 with a total of 3 iterations. The best performance of the distance method in the K-means Clustering process in the Gini ratio dataset is the Manhattan Distance method.

Keywords: K-means; Euclidean Distance; Manhattan Distance; Gini Ratio.

DAFTAR ISI

DAFTAR TABEL

HALAMAN

DAFTAR GAMBAR

HALAMAN

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kondisi ekonomi yang beragam di Indonesia membuat penyeragaman pemerataan ekonomi perlu diukur, termasuk di Sumatera Utara. Pengukuran pemerataan ekonomi dapat dilihat melalui rasio Gini yang umumnya digunakan sebagai alat ukur distribusi pendapatan atau kekayaan pada suatu populasi agar merata. Semakin tinggi rasio Gini maka semakin tinggi pula Tingkat ketidakmerataannya. (Hamdani & Mayshelly, 2023)

Menganalisis *distance* dalam suatu populasi kelompok sangatlah penting demi melihat kemiripan antar isi dari populasi tersebut termasuk dalam hal persebaran rasio Gini, agar dapat melihat *distance* tersebut maka perlu digunakan metode untuk menganalisisnya. Metode Euclidean dan Manhattan dipilih sebab dengan memanfaatkan metode yang paling akurat, maka kebijakan yang paling efektif dan tepat sasaran dapat diterapkan. Pemilihan metode Manhattan dan Euclidean sebagai metode untuk diuji dan dibandingkan disebabkan keduanya mudah, hemat waktu , prosesnya cepat dalam menganalisis data *distance,* dan cukup populer dikalangan peneliti sehingga sudah sepatutnya diuji untuk melihat mana yang terbaik. (Solikhun, 2022)

Pada penelitian yang dilakukan oleh David Wijaya dengan menguji kedua metode untuk mengidentifikasi jenis batik, dinyatakan bahwa analisis yang dilakukan oleh Manhattan Distance cendrung lebih unggul ketimbang hasil analisis yang dilakukan pada Euclidean Distance(Wijaya & Widiarti, 2024). Sedangkan

pada penelitian yang dilakukan oleh Lu'luul Maknun terhadap pendeteksian dini Covid-19 menyatakan bahwasannya analisis yang dilakukan menunjukkan bahwasannya Euclidean Distance lebih unggul dalam hal mengukur jarak ketimbang Manhattan Distance (Maknun, Syukur, Affandy, & Soeleman, 2022)

Adapun perbedaan dari kedua metode yang hendak diuji terletak di sejumlah aspek misalnya saja di metode Manhattan Distance, analisis jarak yang dilakukan akan menempuh proses pengukuran koordinat dengan skema atas-bawah dan kiri-kanan sehingga konsepnya seperti mengukur jarak suatu tempat ke tempat lain dengan melalui jalanan di perkotaan yang berbentuk blok. Euclidean sendiri melakukan analisis untuk mengukur jarak dengan cara mengukur koordinat tertentu ke koordinat tujuan dengan cara menarik garis lurus tanpa berbelok ,tidak seperti pada Manhattan, Euclidean seperti ketika pesawat melintas maka pesawat itu akan terbang lurus tanpa berbelok tidak seperti ketika mengendarai mobil diperkotaan. (Ghazal et al., 2021)

Distance sendiri dapat dipahami sebagai upaya untuk mengelompokkan anggota cluster berdasarkan kemiripan sifatnya, misalnya saja didalam *cluster* keju kuning terdapat keju kuning yang agak kehijauan dan ada yang agak kebiruan, dengan memanfaatkan *distance* kita dapat mengukur suatu keju akan masuk kelompok keju kuning agak kebiruan atau agak kehijauan dalam *cluster* keju kuning tersebut.

Penelitian ini mengusulkan perbandingan dua metode jarak / *distance*, yaitu metode Manhattan dan Euclidean, dalam analisis data rasio Gini Sumatera Utara menggunakan algoritma K-means. Algoritma K-means digunakan untuk mengelompokkan daerah-daerah dengan tingkat ketidakmerataan yang serupa.

Pemilihan metode jarak merupakan langkah kritis dalam proses K-means,(Siahaan, 2022) dan perbandingan antara metode Manhattan dan Euclidean dapat memberikan pemahaman lebih mendalam tentang cara kedua metode tersebut berperan dalam menghasilkan kelompok-kelompok populasi yang relevan dan sesuai dalam artian menghasilkan data analisis yang lebih baik(Hartono, Eniyati, & Hadiono, 2023).

Dengan demikian, Penelitian ini bertujuan untuk mendalami perbandingan performa algoritma K-means ketika menggunakan metode *distance* Euclidean dan Manhattan dengan memasukkan data rasio GINI sebagai bahan uji. Dengan pemahaman yang lebih mendalam terhadap kinerja kedua metode *distance* tersebut, diharapkan penelitian ini dapat memberikan panduan lebih baik dalam pemilihan metode *distance* yang sesuai untuk kasus sejenis.

1.2. Rumusan Masalah

Adapun rumusan masalah pada penelitian ini meliputi beberapa poin yakni:

1. Bagaimana perbandingan kinerja Metode Manhattan dan Metode Euclidean dalam analisis data rasio Gini Sumatera Utara menggunakan algoritma K-means?
2. Apa terdapat perbedaan signifikan antara hasil distance yang diperoleh dari metode Manhattan dan Euclidean terkait pemahaman pola ketidakmerataan ekonomi di Sumatera Utara melalui proksi Rasio Gini?
3. Apa metode yang lebih unggul dalam mengukur distance ketika menghitung jarak data rasio Gini kabupaten / kota di Sumatera Utara?

1.3. Batasan Masalah

Adapun batasan masalah pada penelitian ini meliputi beberapa poin yakni:

1. Metode yang digunakan ialah Manhattan & Euclidean *Distance* serta algoritmanya berupa K-Means.
2. Dataset yang digunakan meliputi data rasio Gini pada kabupaten / kota di Sumatera Utara dari tahun 2000-2023.
3. Bahasa pemograman yang digunakan yakni Bahasa Python yang diimplementasikan kedalam Google Colabolatory.
4. Aplikasi yang digunakan untuk mengolah data dan menganalisisnya yakni Google Colabolatory.

1.4. Tujuan Penelitian

Didasari dengan pernyataan pada rumusan masalah maka dapat dibangun tujuan penelitian yang terdiri dari beberapa poin, yakni:

1. Mengetahui perbandingan kinerja Metode Manhattan dan Metode Euclidean dalam analisis data rasio Gini Sumatera Utara menggunakan algoritma K-means.
2. Mengetahui perbedaan signifikan antara hasil distance yang diperoleh dari metode Manhattan dan Euclidean terkait pemahaman pola ketidakmerataan ekonomi di Sumatera Utara melalui proksi Rasio Gini.
3. Mengetahui metode yang lebih unggul dalam mengukur distance ketika menghitung jarak data rasio Gini kabupaten / kota di Sumatera Utara.

1.5. Manfaat Penelitian

Adapun manfaat dari penelitian yang dilakukan saat ini yakni diantaranya ialah:

1.5.1. Bagi Peneliti

Bagi Peneliti, penelitian kali ini akan memberikan manfaat dalam menambah wawasan dan pengetahuan terkait kemampuan dan performa dari Metode Manhattan dan Euclidean dalam mengukur *distance*.

1.5.2. Bagi Peneliti Selanjutnya

Bagi peneliti selanjutnya, penelitian ini harapannya dapat menjadi referensi dalam melakukan tindak pengembangan penelitian di masa yang akan datang, baik sebagai panduan dan pedoman penelitian maupun sebagai bahan banding pada penelitian yang membahas *distance* dan *data mining*.

1.5.3. Bagi Pembaca

Bagi pembaca, penelitian ini harapannya dapat digunakan sebagai sarana penambah wawasan terkait kemampuan dan performa masing-masing dari metode Manhattan dan Euclidean dalam mengukur *distance*.

1.5.4. Bagi Pemerintah

Bagi pemerintah, penelitian ini harapannya dapat digunakan sebagai referensi dan bahan pendukung dalam membentuk kebijakan dan keputusan dalam hal memeratakan ekonomi kawasan kabupaten / kota di Sumatera Utara dan dapat menyeimbangkan kesenjangan ekonomi yang ada demi menyeragamkan setiap wilayah.

BAB II

LANDASAN TEORI

2.1. Data Mining

Data mining ialah proses yang dilakukan untuk mengelola dan menganalisis data yang ada dengan tujuan menemukan pola-pola tertentu dan dapat menghasilkan penafsiran-penafsiran tertentu. Data Mining merupakan salah satu tahap kunci dalam proses *Knowledge Discovery in Databases* (KDD), sebuah pendekatan sistematis untuk menemukan pengetahuan yang berharga dari kumpulan data. Proses KDD terdiri dari beberapan tahapan. Tahap utama dalam proses KDD adalah data mining, di mana teknik-teknik analisis statistik dan komputasi digunakan untuk menemukan pola, hubungan, atau informasi yang tersembunyi dalam data. Data mining bisa dilakukan dengan beragam metode seperti regresi, *clustering*, klasifikasi, atau *association rule mining* yang penggunaannya disesuaikan dengan kebutuhan. (Zai, 2022)

Data Mining memiliki pengaplikasian luas di berbagai bidang, termasuk bisnis, keuangan, kesehatan, pemasaran, dan ilmu pengetahuan lainnya. Dalam konteks bisnis, data mining dapat dimanfaatkan untuk memprediksi penjualan dan segmentasi pelanggan, pada konteks lain dapat bermanfaat pula untuk beragam hal seperti mendeteksi kecurangan, pengembangan obat baru, mengukur persebaran penyakit dan dsb. Dengan kemampuannya untuk mengeksplorasi dan menganalisis data secara mendalam, Data Mining memberikan nilai tambah yang signifikan bagi organisasi dalam pengambilan keputusan yang lebih baik, peningkatan efisiensi operasional, dan pengembangan kebijakan yang lebih efektif.(Anjumi et al., 2022)

2.2. Distance

Jarak atau lebih dikenal sebagai *Distance,* dalam data mining konteks *distance* memiliki makna sebagai metode pengukuran yang dapat digunakan untuk melihat kedekatan antara dua atau lebih objek atau titik dalam ruang multidimensi. Konsep jarak / *distance* sangatlah penting dalam melakukan data mining yang berfokus ke upaya klasifikasi, clustering, dan pencarian pola-pola tertentu. Pengukuran jarak memiliki dampak signifikan dalam berbagai algoritma data mining. Sebagai contoh, dalam klastering, algoritma seperti K-means mengelompokkan data berdasarkan kedekatan antara titik-titiknya, yang ditentukan oleh jarak antara mereka. Selain itu, konsep jarak juga diterapkan dalam pengenalan pola, pencarian rekaman terdekat dalam basis data spasial, pengolahan citra, dan bidang-bidang lain di luar data mining. Dalam konteks ini, metode pengukuran jarak dapat beragam tergantung pada kebutuhan dan sifat data yang dihadapi. (Agus, 2023)

Konsep *distance* dapat diterapkan pada berbagai jenis data, mulai dari data numerik hingga data kategorikal. Dalam data numerik, seperti data yang terdiri dari bilangan riil atau bilangan bulat, jarak sering diukur dengan Euclidean. Dalam Euclidean, jarak antara dua titik dihitung sebagai garis lurus yang menghubungkan kedua titik tersebut. Ini dikenal sebagai *Euclidean distance*, yang dapat dihitung menggunakan rumus Pythagoras. Analisis pada data kategorikal yang mana datanya tidak memiliki nilai numerik langsung, metode pengukuran *distance* yang lain akan lebih efektif. *Hamming distance* sering digunakan dalam konteks data katagorikal. *Hamming distance* mengukur jumlah posisi di mana dua vektor mengalami perbedaan. Ini sangat berguna ketika bekerja dengan data biner atau kategorikal, di

mana peneliti hanya tertarik pada perbedaan dalam penentuan kategori atau nilai.(Rahayu, Fauzan, & Harliana, 2022)

2.3. Algoritma K-Means

Algoritma K-means merupakan algoritma yang umum digunakan dalam upaya melakukan data mining. Tujuan dan manfaat dari diterapkannya algoritma K-means yakni untuk mengkelompokkan data-data dalam suatu golongan tertentu *(clustering)*. Algoritma K-means meliputi beberapa tahapan yang diantaranya terdapat pengukuran jarak disetiap titik data terhadap pusat inti cluster, pengelompokan titik-titik data ke cluster terdekat, hingga ke upaya pembaharuan pusat cluster berdasarkan rata-rata titik dalam setiap cluster. Seluruh proses dalam algoritma K-Means akan terus diulang hingga tidak ada perubahan yang signifikan dan iterasi yang ditetapkan telah tercapai. Hasil analisis dapat digunakan untuk beragam kepentingan misalnya seperti segmentasi pasar, analisis citra / grafik, maupun kegunaan lainnya. (Kurniawan, Hasibuan, & Hasibuan, 2023)

K-means sangat kerap digunakan dalam penelitian namun nyatanya terdapat beberapa kelemahan dibalik kepopulerannya, diantara kelemahan dari K-means terdapat kelemahan terkait pusat inti cluster, penentuan titik pusat cluster yang secara acak diawal dapat berdampak pada hasil cluster. Kelemahan lainnya, K-means akan menanggap setiap data memiliki sifat yang sama dalam hal ukuran, kepadatan, dan bentuk geometris data sehingga data cluster yang dihasilkan berpotensi menjadi tidak maksimal. (Mirantika, Syamfithriani, & Trisudarmo, 2023)

2.4. Manhattan Distance

Manhattan Distance, yang juga dikenal sebagai jarak kota, adalah salah satu metode yang umum digunakan dalam analisis data mining untuk clustering dan penemuan pola tertentu. Metode ini bekerja dengan mengukur jarak antara dua titik dengan menghitung perbedaan koordinat mereka. Keunggulan utama Manhattan Distance adalah kemampuannya untuk menangani data yang memiliki dimensi yang tidak seragam, serta kemampuannya mempertahankan ketepatan dalam pengukuran jarak dalam situasi di mana garis lurus tidak dapat diterapkan, seperti dalam kasus data spasial atau data dengan pola yang tidak teratur. Meskipun Manhattan Distance sering kali dianggap sebagai pilihan yang bagus, namun tidak selalu menjadi solusi terbaik untuk setiap kasus. Seperti halnya metode jarak lainnya, keefektifan Manhattan Distance sangat tergantung pada karakteristik dan tujuan analisis dari data yang sedang diteliti. Dalam beberapa kasus, terutama ketika data terdistribusi secara berbeda-beda, metode lain mungkin menjadi lebih efektif daripada Manhattan Distance. Mempertimbangkan secara hati-hati pilihan metode jarak yang sesuai dengan karakteristik data dan tujuan analisis akan menjadi hal yang krusial. Dengan memahami kekuatan dan keterbatasan setiap metode, serta sifat-sifat khusus dari dataset yang sedang dianalisis, kita dapat memilih pendekatan yang paling sesuai untuk mencapai tujuan analisis data dengan maksimal.(Mughnyanti & Hafiz Nanda Ginting, 2023)

2.5. Euclidean Distance

Euclidean distance adalah salah satu metode pengukuran jarak yang paling umum digunakan dalam berbagai bidang analisis data yang diantaranya meliputi klastering, klasifikasi, dan pencarian pola. Konsepnya didasarkan pada geometri Euclidean, dimana jarak antara dua titik diukur sebagai garis lurus yang menghubungkan kedua titik tersebut. Dalam menghitung Euclidean Distance, rumus Phythagoras dapat digunakan dalam kalkulasinya. Keunggulan utama Euclidean distance adalah kemampuannya untuk memberikan gambaran geometris tentang kedekatan antara titik-titik data, serta kemampuannya untuk menangani data dengan dimensi yang sama atau berbeda. Namun, Euclidean distance juga memiliki beberapa kelemahan, seperti misalnya sensitif terhadap perbedaan skala antara dimensi, di mana dimensi dengan rentang nilai yang lebih besar dapat mendominasi perhitungan jarak. Selain itu, Euclidean distance nyatanya tidak selalu efektif dalam menangani data dengan pola yang tidak teratur sehingga kondisi data menjadi penentu dalam penerapannya pula. Dengan demikian, dalam menggunakan Euclidean distance peneliti perlu mempertimbangkan karakteristik data yang dihadapi dan kadang-kadang diperlukan teknik tambahan dalam melakukan analisis dengan metode, seperti normalisasi data atau penggunaan metode pengukuran jarak yang lebih sesuai dengan struktur data yang ada meskipun pengunaannya yang mudah dan digemari banyak peneliti dalam mengolah dan menganalisis data mining.(Andika Naufal, Setiadi, & Trisnawijayana, 2022)

2.6. Badan Pusat Statistik (BPS)

Badan Pusat Statistik (BPS) adalah lembaga pemerintah di Indonesia yang memiliki tugas untuk melakukan pengumpulan, pengolahan, analisis, dan publikasi data statistik yang berkaitan dengan kondisi sosial, ekonomi, dan demografi negara. BPS memiliki peranan yang penting dalam menyediakan informasi yang akurat dan andal bagi pemerintah, sektor bisnis, pendidikan, organisasi masyarakat, dan masyarakat umum untuk mendukung pengambilan keputusan yang berbasis data. BPS mengumpulkan data dari berbagai sumber, termasuk survei yang dilakukan secara berkala, sensus penduduk dan ekonomi, serta data administratif dari berbagai instansi pemerintah. Data yang dikumpulkan oleh BPS mencakup berbagai aspek kehidupan, mulai dari pendapatan dan kemiskinan, tenaga kerja dan lapangan pekerjaan, inflasi dan harga konsumen, pendidikan dan kesehatan, hingga data geografis dan lingkungan, serta data spesifik seperti rasio Gini. Selain itu, BPS juga bertanggung jawab atas penelitian dan pengembangan metodologi statistik, standar pengukuran, dan pengelolaan sistem informasi statistik nasional. Dengan memberikan akses terhadap data statistik yang lengkap dan dapat dipercaya, BPS berperan penting dalam memfasilitasi pembangunan nasional, perencanaan kebijakan, evaluasi program, serta meningkatkan transparansi dan akuntabilitas pemerintah. Selain itu, BPS juga berperan dalam memfasilitasi kerja sama internasional dalam pertukaran data statistik untuk mendukung pembangunan berkelanjutan dan pencapaian tujuan pembangunan global. (Lani, Hendra, & Khalid, 2023)

Dalam penelitian ini, BPS berperan sebagai sumber data rasio Gini yang akan diuji dalam menentukan metode distance mana yang terbaik antara Manhattan dan

Euclidean. Dengan demikian data yang akan digunakan dalam penelitian ini ialah data sekunder yang berasal dari BPS

2.7. Rasio Gini

Rasio Gini adalah alat ukur yang bisa digunakan dalam mengukur ketidaksetaraan atau distribusi pendapatan di dalam suatu populasi atau wilayah. Rasio ini berfungsi sebagai alat untuk mengukur sejauh mana pendapatan atau kekayaan diantara individu atau kelompok dalam masyarakat terdistribusi secara merata. Rasio Gini memiliki rentang nilai antara 0 hingga 1, dimana nilai 0 menunjukkan distribusi pendapatan yang sempurna (setiap individu memiliki pendapatan yang sama), sementara nilai 1 menunjukkan distribusi yang paling tidak merata (semua pendapatan dikonsentrasikan pada satu individu). Data untuk menghitung Rasio Gini dapat diperoleh dari berbagai sumber, termasuk survei pendapatan rumah tangga, data pajak, atau data administratif lainnya.

Metode perhitungan Rasio Gini meliputi pengurutan individu atau rumah tangga berdasarkan pendapatan mereka, lalu menghitung proporsi kumulatif dari total pendapatan yang diterima oleh populasi. Perbedaan antara garis kumulatif pendapatan yang seimbang (garis Lorenz) dan garis kumulatif pendapatan aktual digunakan untuk menghitung Rasio Gini. Meskipun terdengar efektif , namun Rasio Gini masih memiliki kekurangan sehingga penggunaannya harus disesuaikan dengan kebutuhan. Sebagai contoh, Rasio Gini tidak memberikan informasi tentang sumber atau penyebab ketidaksetaraan dan hanya memberikan gambaran umum terkait tingkat ketidaksetaraan ekonomi dalam suatu populasi. Dikarenakan alasan tersebut, Rasio Gini kerap digunakan bersama-sama dengan

indikator lainnya untuk memberikan pemahaman yang lebih lengkap tentang ketidaksetaraan dan untuk merumuskan kebijakan yang bertujuan untuk mengurangi ketidaksetaraan ekonomi dan meningkatkan kesejahteraan sosial dikawasan.(El Islami & Fitrianto, 2023)

2.8. Google Colaboratory (Menggunakan Bahasa Python)

Google Colaboratory (Colab) adalah platform cloud yang disediakan secara gratis oleh Google untuk menjalankan kode Python. Dengan kemudahan akses dan fleksibilitasnya, Colab kerap digunakan oleh kalangan ilmuwan data, peneliti, dan pengembang perangkat lunak. Dengan menggunakan Colab, peneliti dapat menulis, menjalankan, dan berbagi kode Python tanpa perlu menginstal atau mengkonfigurasi perangkat lunak di komputer yang digunakan.

Dengan memanfaatkan Colab, peneliti dapat melakukan pemrosesan data besar, pelatihan model kecerdasan buatan yang kompleks, dan eksperimen komputasi lainnya dengan cepat dan efisien. Pengguna dapat menyimpan notebook Colab secara langsung di Google Drive mereka dan berbagi dengan rekan tim atau pembimbing, hal ini memungkinkan kolaborasi secara real-time, di mana beberapa pengguna dapat bekerja bersama pada notebook yang sama, mempercepat proses pengembangan dan meningkatkan produktivitas tim. Dengan semua fitur dan kemudahan yang ditawarkannya, Colab telah menjadi alat yang sangat berharga dalam penelitian dan pengembangan proyek-proyek berbasis Bahasa Python, terutama dalam konteks skripsi di mana eksplorasi data, dan analisis statistik seringkali menjadi bagian penting dari proses penelitian. Dengan menggunakan

Colab, peneliti dapat lebih fokus pada eksplorasi ide dan konsep, tanpa harus khawatir tentang infrastruktur teknis yang rumit.

Penelitian ini akan menggunakan Google Colab dengan Bahasa Python sebab pengaksesan yang mudah dan kemampuan Google Colab tersebut dalam mengelola pengujian analisis distance.

2.9. Penelitian Terdahulu

Adapun penelitian terdahulu yang digunakan sebagai referensi peneliti dalam melakukan penelitian ini yakni sebagai berikut:

Tabel 2.1. Penelitian Terdahulu

Tahun	Nama Peneliti	Judul	Hasil
2023	Suraya, Muhammad Sholeh, dan Dina Andayati	Comparison of distance metric in k-mean algorithm for clustering wheat grain datasheet	Data yang diolah dengan menggunakan Manhattan distance cendrung memiliki gambaran pengelompokkan clustering data pada cluster 0 dan sedikit pada cluster 1, sedangkan data yang diolah dengan Euclidean memiliki kecendrungan cluster 1 mendominasi lebih banyak data ketimbang cluster 0.
2024	David Wijaya, dan Anastasia Rita Widiarti	Batik classification using KNN algorithm and GLCM features extraction	Analisis yang dilakukan oleh Manhattan Distance cendrung lebih unggul ketimbang hasil analisis yang dilakukan pada Euclidean Distance, meski demikian diperlukan penelitian lebih lanjut sebab data yang digunakan pada penelitian banyak yang tidak terbaca.
2022	Lu'luul Maknun, Abdul Syukur, Affandy, dan Moch Arief Soeleman	Deteksi Dini Covid-19 Melalui Citra Ct-Scan Paru-Paru Menggunakan K-Nearest Neighbor Dengan Komparasi Jarak	Analisis yang dilakukan menunjukkan bahwasannya Euclidean Distance lebih unggul dalam hal mengukur jarak ketimbang Manhattan Distance dalam kasus

			pengukuran jarak pada pendeteksian dini Covid-19.
2020	Rozzi Kesuma Dinata, Hafizal Akbar, dan Novia Hasdyna	Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus	Analisis yang telah dilakukan menunjukkan bahwasannya analisis Manhattan Distance terlihat lebih unggul dibandingkan dengan analisis yang dilakukan menggunakan Euclidean Distance dalam klasifikasi transportasi bus.
2022	Mohamad Sugeng Pangestu, dan Maulida Ayu Fitriani	Perbandingan Perhitungan Jarak Euclidean Distance, dan Manhattan Distance, dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means	Hasil analisis menunjukkan bahwasannya Euclidean Distance lebih unggul ketimbang Manhattan Distance dalam mengukur jarak pada data bibit padi apabila menggunakan algoritma K-means.

Sumber: (Dinata, Akbar, & Hasdyna, 2020; Maknun et al., 2022; Pangestu & Fitriani, 2022; Suraya, Sholeh, & Andayati, 2023; Wijaya & Widiarti, 2024)

BAB III

METODOLOGI PENELITIAN

3.1. Jenis Penelitian

Jenis penelitian adalah klasifikasi berbagai metode penelitian ilmiah, termasuk deskriptif, eksperimental, kuantitatif, kualitatif, korelasional, dan gabungan yang mana nantinya akan dipadukan dengan pola berpikiran terbuka dan menyeluruh yang memperhitungkan semua fakta.(Yudha, Widodo, & Febiharsa, 2023)

Dalam penelitian ini, jenis penelitian yang akan diterapkan ialah penelitian kuantitatif dengan melakukan analisis komparatif terhadap performa metode Manhattan Distance dan Euclidean Distance dalam mengelola data rasio Gini di Sumatera Utara untuk melihat distance yang terdapat didalamnya.

3.2. Definisi Operasional

Definisi operasional adalah cara untuk mengukur atau mendefinisikan suatu konsep atau variabel dalam konteks penelitian dengan tujuan memberikan panduan konkret tentang bagaimana konsep pada penelitian akan diukur atau diamati dalam pelaksanaan, sehingga memungkinkan peneliti untuk mengumpulkan data yang konsisten dan dapat diandalkan untuk analisis lebih lanjut.(Hasugian & Wardarita, 2022)

Agar memudahkan pemahaman dalam penelitian ini, maka dapat dibentuk sebuah definisi operasional sebagai berikut:

Tabel 3.1. Definisi Operasional Variabel

Variabel	Definisi
Manhattan Distance	Manhattan distance adalah metode pengukuran jarak antara dua titik dalam ruang dengan menjumlahkan perbedaan absolut dari koordinat x dan y.
Euclidean Distance	Euclidean distance adalah metode pengukuran jarak antara dua titik dalam ruang dengan menggunakan teorema Pythagoras untuk menghitung jarak lurus antara titik-titik tersebut.
Distance	Distance adalah pengukuran jarak antara titik-titik data, digunakan untuk menentukan seberapa dekat atau jauh setiap titik dari yang lain, memungkinkan pembentukan cluster yang optimal.

Sumber: Dokumentasi Peneliti, 2024

3.3. Teknik Pengambilan Sampel

Convenience Sampling adalah metode pengambilan sampel tanpa pemilihan maupun penyaringan data sesuai kriteria-kriteria tertentu secara formal melainkan dikarenakan data tersebut mudah diakses dan tersedia untuk publik. (Djouzi, Beghdad-bey, & Amamra, 2023)

Pada penelitian ini, peneliti akan menggunakan teknik pengambilan sampel yakni *Convenience Sampling* yang mana data tersebut diambil dari Badan Pusat Statistik (BPS) Sumut.

3.4. Teknik Pengumpulan Data

Teknik pengumpulan data adalah pendekatan atau metode yang digunakan untuk mengumpulkan informasi atau data dari subjek penelitian. Ini melibatkan strategi dan prosedur tertentu yang digunakan untuk mendapatkan data yang relevan dan representatif sesuai dengan tujuan penelitian.(H et al., 2023)

Pada penelitian ini, Teknik pengumpulan data yang digunakan ialah:

1. Studi Literatur

   Studi literatur adalah teknik pengumpulan data dimana peneliti mengkaji dokumen seperti laporan maupun catatan yang relevan dengan penelitian yang dilakukan. Pada penelitian ini, peneliti melakukan studi literatur terhadap data Rasio Gini yang terdapat di situs resmi BPS.

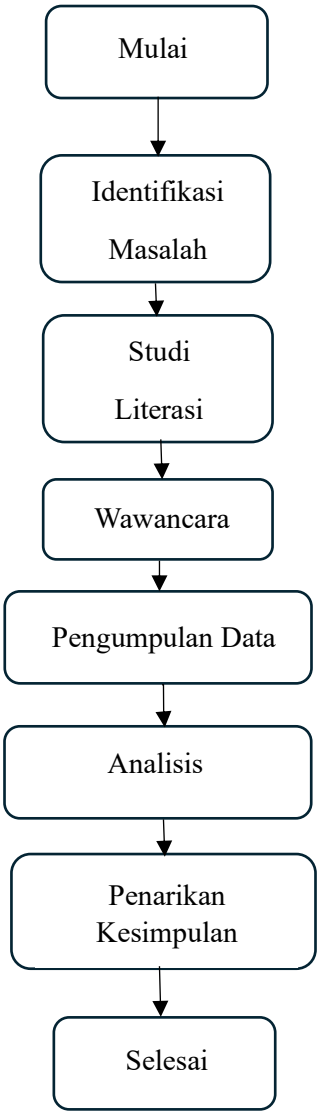
2. Wawancara

   Wawancara adalah teknik pengumpulan data dimana peneliti melakukan interaksi langsung dengan sejumlah narasumber untuk memperoleh pengetahuan dan informasi yang dibutuhkan. Pada penelitian ini, peneliti melakukan wawancara dengan petugas BPS yang terdapat di pojok statistic di Universitas Muhammadiyah Sumatera Utara (UMSU) untuk mengetahui terkait tentang gambaran apa itu rasio Gini dan mengapa diperlukan pengukuran distance didalamnya.

3.5. Tahapan Penelitian

Tahapan penelitian adalah serangkaian langkah sistematis yang dilakukan oleh seorang peneliti untuk mengidentifikasi masalah penelitian dan merancang penelitian. Proses ini bertujuan untuk mencapai tujuan penelitian yang telah

ditetapkan dengan cara yang terstruktur dan terorganisir. Penelitian ini memiliki tahapan penelitian sebagai berikut:



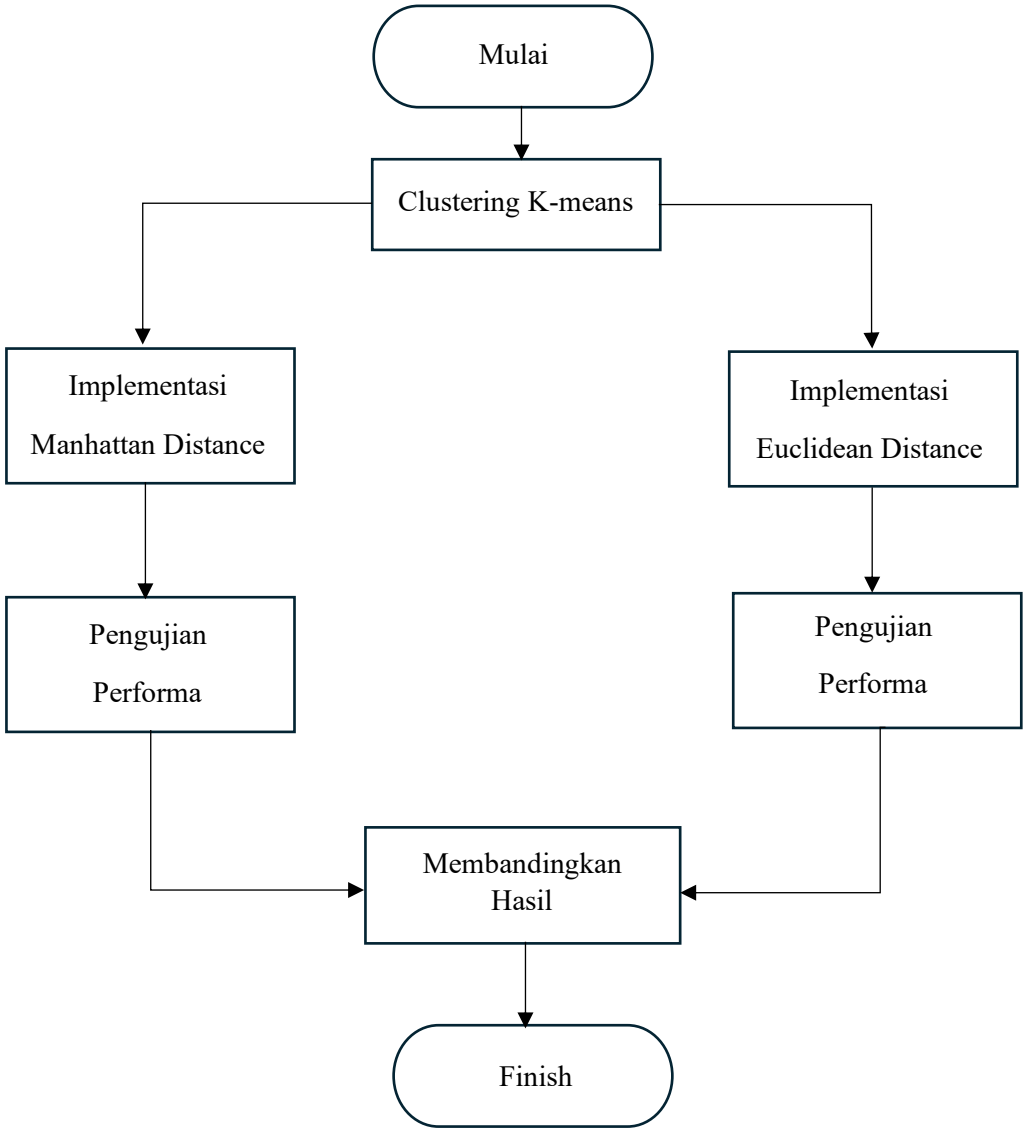
Sumber: Dokumentasi Peneliti, 2024

Gambar 3.1. Alur Tahapan Penelitian

3.6. Teknik Analisis

Penelitian ini akan melakukan pengujian pada 2 metode *distance* yakni Manhattan Distance dan Euclidean Distance untuk melihat metode mana yang lebih baik dalam mengelolah data *cluster* yang berasal dari pengklasteran rasio Gini dengan

membandingkan 2 indikator pada hasil analisis yakni waktu iterasi, dan Dunn Index. Adapun gambaran dari proses analisis yang akan dilakukan yakni sebagai berikut:



Sumber: Dokumentasi Peneliti, 2024

Gambar 3.2. Alur Proses Analisis

3.7. Jadwal dan Waktu Penelitian

Penelitian ini mengambil data yang akan digunakan sebagai penguji antara 2 metode berasal dari Badan Pusat Statistik (BPS) Sumatera Utara yang beralamatkan di Jl. Asrama No.179, Kota Medan, Sumatera Utara.

Adapun waktu penelitian pada penelitian ini akan mencangkup jadwal sebagai berikut:

Tabel 3.2. Aktivitas dan Waktu Penelitian

No.	Aktivitas Penelitian	Bulan			
		Januari	Febuari	Maret	April
1.	Penelitian Pendahuluan (Prariset)				
2.	Penyusunan Proposal				
3.	Pembimbingan Proposal				
4.	Seminar Proposal				
5.	Perbaikan Hasil revisi				
6.	Pengumpulan Data				
7.	Pengolahan dan Analisis Data				
8.	Penyusunan Skripsi (Laporan Penelitian)				
9.	Pembimbingan Skripsi				
10.	Sidang Meja Hijau				
11.	Penyempurnaan Skripsi dan Penulisan Artikel Jurnal				

Sumber: Dokumentasi Peneliti, 2024

3.8. Analisis K-Means dengan Manhattan Distance

Dalam melakukan pengklasteran pada data mining, proses perhitungan jarak / *distance* menjadi sangat penting untuk mengelompokkan data-data yang ada dengan melihat kemiripan pada setiap data. Manhattan Distance adalah proses perhitungan distance yang menhitung jarak suatu data ke titik pusat cluster dengan menghitung jarak berdasarkan kordinat (bukan menarik garis lurus). Adapun rumus dari Manhattan Distance ialah:

$$d(x, y) = \sum |x - y| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

Dimana,

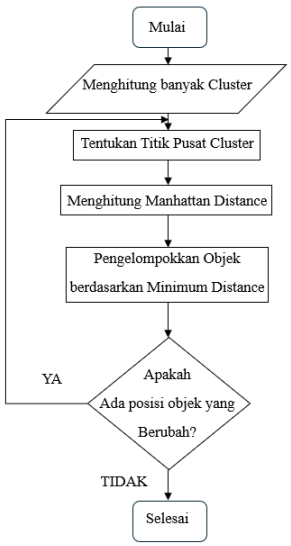
d = Distance

x = Koordinat Longtitude (sumbu x)

y = Koordinat Latitude (sumbu y)

Sumber: (Eviana, Fauzan, Harliana, & Putra, 2022)

Adapun proses pengklasteran K-Means yang menggunakan Manhattan Distance sebagai perhitungan jaraknya dapat sebagai berikut:



Sumber: Dokumentasi Peneliti, 2024

Gambar 3.3. Diagram Alur K-Means Dengan Manhattan Distance

Berdasarkan began diagram alur diatas, bisa kita pahami bahwasannya proses pengklasteran dimulai dengan menghitung banyak klaster yang akan dibuat, setelahnya proses penentuan titik pusat pada setiap kluster (Sriadhi, Gultom, & Martiano, 2020). Proses berlanjut dengan mulai menghitung data-data yang hendak diolah distancenya sehingga dapat diketahui bagaimana pengklasterannya berdasarkan jarak data sampai ke titik pusat dengan menggunakan Manhattan Distance dimana jaraknya diukur dengan mengukur jarak koordinat layaknya berjalan di perkotaan padat (berdasarkan belokan pada blok koordinat), setelah data jarak / distance diketahui maka dapat dilakukan pengelompokkan klasternya. Tahapan berikutnya ialah melihat apakah ada posisi data yang berubah apabila dibandingkan dengan pengolahan sebelumnya, apabila ada posisi data yang berubah maka proses kembali diulang ke tahap awal dan terus berlanjut hingga akhirnya tidak ada posisi yang berubah lagi sehingga pengklasteran dapat dikatakan selesai.

3.9. Analisis K-Means dengan Euclidean Distance

Perhitungan klastering tidak akan terlepas dari perhitungan jarak untuk menentukan anggota dari klaster tersebut, selain metode Manhattan Distance terdapat juga metode lain yang kerap digunakan oleh para peneliti yakni Euclidean Distance, dalam Euclidean Distance yang ditemukan 2300 tahun lalu oleh Euclid, perhitungan jarak suatu data ke pusat klaster dihitung dengan menarik garis lurus. Adapun rumus dari Euclidean Distance yakni sebagai berikut:

$$d(x, y) = \sqrt{\sum}(x - y)\ 2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

Dimana,

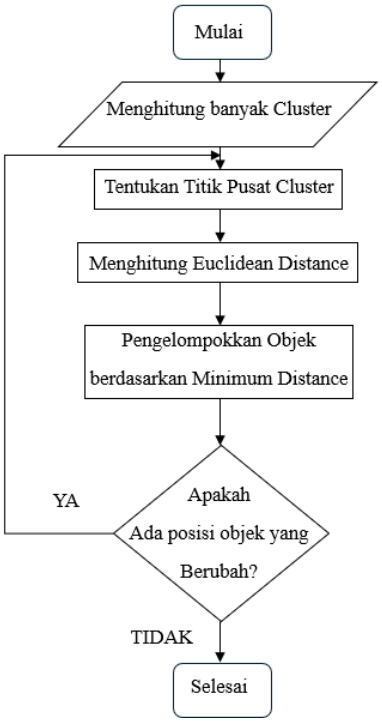
d = Distance

x = Koordinat Longtitude (sumbu x)

y = Koordinat Latitude (sumbu y)

Sumber: (Eviana et al., 2022)

Adapun proses pengerjaan klastering K-means yang menggunakan Euclidean Distance dalam menghitung jarak data-data yakni sebagai berikut:



Sumber: Dokumentasi Peneliti, 2024

Gambar 3.4. Diagram Alur K-Means Dengan Euclidean Distance

Proses pengklasteran dimulai dengan menghitung banyak klaster, di ikuti dengan penentuan titik pusat setiap klaster setelahnya (Martiano, Sari, & Akbar, 2023) dan berlanjut dengan perhitungan jarak data-data yang ada ke titik pusat klaster, perhitungan jarak dilakukan dengan cara menarik garis lurus dari posisi data ke titik pusat klaster layaknya menarik garis menggunakan penggaris. Apabila penghitungan telah dilakukan maka akan dilakukan pengelompokkan berdasarkan

kedekatan jarak antara posisi data dengan titik pusat klaster, jika ada data yang berubah apabila dibandingkan dengan data sebelumnya, maka dilakukan pengulangan ketahap awal yang berlanjut hingga ke tahap dimana data dibandingkan dengan hasil pengklasteran sebelumnya. Apabila tidak ada perubahan lagi maka proses pengklasteran dapat dikatakan sudah selesai.

3.10. Analisis Perbandingan Waktu Iterasi

Iterasi adalah pengulangan yang dilakukan pada data mining dengan tujuan memperoleh data yang stabil dan optimal setelah tercapainya konvergensi (kondisi dimana tidak ada lagi perubahan yang signifikan sehingga dapat dianggap data olahan telah mencapai bentuknya yang baik). (Siregar & Octariadi, 2021)

Dalam melakukan analisis waktu iterasi pada data mining, proses dimulai dengan melakukan pengolahan pengklasteran data. Proses ditempuh dengan terus mengulangi proses olah data hingga hasil olahannya mencapai titik konvergensi dimana tidak ada lagi perubahan yang signifikan apabila dibandingkan dengan hasil pengolahan sebelumnya. Pada penelitian ini, pengklasteran dengan Manhattan Distance dan Euclidean Distance akan dihitung berapa lama iterasi yang dibutuhkan oleh setiap metode hingga akhirnya mencapai titik kondisi dimana konvergensi diperoleh, semakin cepat waktu iterasinya maka akan semakin baik dan begitu pula sebaliknya apabila semakin lama.

3.11. Analisis Perbandingan Dunn Index

Dunn Index adalah indeks yang digunakan untuk melihat kemiripan data-data yang ada pada kluster, semakin tinggi nilai Dunn Index pada suatu metode maka semakin

baik pula metode itu dalam melakukan pengklasteran.(Septianingsih, 2022) Adapun rumus untuk menghitung Dunn Index yakni sebagai berikut:

$$Dunn\ Index = \frac{Dmin}{Dmax} \text{.....(3.3)}$$

Dimana,

Dmin = Jarak terdekat antara dua klaster berbeda

Dmax = Jarak terjauh antar dua titik dalam satu klaster

Sumber:(Septianingsih, 2022)

BAB IV

HASIL DAN PEMBAHASAN

4.1. Data Awal

Penelitian ini menggunakan data rasio Gini yang telah disediakan oleh BPS dimana data yang digunakan meliputi data dari tahun 2000 s/d 2023 dimana apabila angka rasio Gini semakin mendekati 0 maka berarti pemerataan pendapatan masyarakat sudah baik / merata, sedangkan apabila mendekati 1 maka artinya pemerataan pendapatan terpusat pada orang-orang tertentu saja yang artinya pemerataannya tidak baik.

Adapun data dari rasio Gini yang akan peneliti olah untuk menguji metode Manhattan atau Euclidean untuk melihat meotde yang paling efektif untuk digunakan pada rasio gini ialah sebagai berikut:

Tabel 4.1. Data Rasio Gini Sumatera Utara 2000-2023

No.	Kabupaten/ Kota	Tahun	Nilai Gini Rasio
1.	Nias	2000	0,2166
2.	Asahan	2000	0,2267
3.	Binjai	2000	0,2586
4.	Medan	2000	0,2791
5.	Tapanuli Selatan	2000	0,2128
6.	Tapanuli Tengah	2000	0,2611
...	....	....	.....
802.	Gunung Sitoli	2023	0,3080

Sumber: BPS, 2023

4.2. Pengujian Sistem Data Mining K-means Clustering Menggunakan Manhattan Distance

4.2.1. Proses Input Library

```
[1]  #Import library Pandas untuk mengolah file yang akan dianalisis
     import pandas as pd
     import seaborn as sns
     from sklearn.cluster import KMeans
     from sklearn.metrics import silhouette_samples, silhouette_score
     import matplotlib.pyplot as plt
     from sklearn.metrics import silhouette_score
     from mpl_toolkits.mplot3d import Axes3D
     import numpy as np
     from sklearn.metrics.pairwise import manhattan_distances
     from sklearn.metrics.pairwise import pairwise_distances_argmin_min
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.1. Kode Input Library

Proses clustering dimulai dengan memanggil library sebagai fungsi untuk menjalankan proses analisis pada file yang akan digunakan, adapun library yang digunakan ialah library Pandas dan Seaborn dengan cara mengimpornya. Penggunaan sklearn.cluster dan sklearn.metrics dimanfaatkan agar proses algoritma K-means dapat dilakukan. Impor berikutnya ialah matplotlib.pyplot yang mana berguna untuk menampilkan grafik koordinat yang akan digunakan sebagai wadah visualisasi clustering. Impor terakhir yakni ialah numpy yang mana tujuannya agar dapat menggunakan sklearn.metrics.pairwise yang bermanfaat untuk melakukan proses *Manhattan Distance.*

4.2.2. Proses Input Data File

```
[2]  # Mengupload file CSV untuk diolah (gini_rasio_a.csv)
     from google.colab import files
     uploaded = files.upload()
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.2. Kode Fitur Input Data

Kode pada tahapan ini digunakan untuk mengunggah file CSV bernama "gini_rasio_a.csv" dari komputer pengguna ke program yang sedang dibangun. Fungsi `files.upload()` dari modul `google.colab` memungkinkan pengguna untuk memilih file dari komputer pengguna dan mengunggahnya ke sistem *Colab* yang sedang berjalan. Ini adalah langkah awal yang sering digunakan dalam proses analisis data untuk mengimpor dataset yang akan dianalisis. Setelah file diunggah, dataset akan tersedia dalam bentuk objek file yang dapat diakses dan dimanipulasi menggunakan library seperti Pandas. Dengan mengunggah file CSV ini, pengguna dapat melakukan berbagai operasi analisis data seperti pemrosesan data mining dan visualisasinya. Proses ini mempermudah pengguna untuk membawa data dari komputer pengguna ke dalam sistem cloud *Google Colab*.

4.2.3. Proses Pengecekan Data

```
#Mengecek isian file yang akan diolah
gini_data = pd.read_csv('gini_rasio_a.csv', delimiter=';', usecols = ['Tahun', 'Gini Rasio'])
gini_data.head()
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.3. Kode Fitur Pengecekan Data

Kode di atas digunakan untuk membaca dan menampilkan isi awal dari file CSV bernama "gini_rasio_a.csv" yang telah diunggah sebelumnya. Library Pandas menyediakan fungsi `read_csv` untuk membaca file CSV, dan dalam hal ini, file

dibaca dengan pemisah delimiter titik koma (`;`). Parameter `usecols` digunakan untuk memilih hanya kolom tertentu, yaitu "Tahun" dan "Gini Rasio", yang relevan untuk analisis yang akan dilakukan. Setelah data dibaca ke dalam DataFrame Pandas yang dinamai `gini_data`, fungsi `head()` dipanggil untuk menampilkan lima baris pertama dari DataFrame tersebut. Ini memungkinkan pengguna untuk memverifikasi bahwa data telah diimpor dengan benar dan untuk melihat sekilas struktur dan konten data yang akan dianalisis. Langkah ini penting untuk memastikan bahwa data sesuai dengan ekspektasi sebelum melakukan analisis lebih lanjut.

4.2.4. Proses Pengecekan Data dengan tabel koordinat

```
#Memunculkan data yang akan diolah kedalam bentuk bagan koordinat, bagan pra-pengelompokkan.
sns.scatterplot(data = gini_data, x = 'Tahun', y = 'Gini Rasio')
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.4. Kode Fitur Pengecekan Data Dengan Tabel Koordinat

Kode di atas digunakan untuk membuat visualisasi data dalam bentuk tabel koordinat scatter plot menggunakan library Seaborn. Fungsi `scatterplot` dari Seaborn membantu memvisualisasikan hubungan antara dua variabel dalam dataset. Dalam penelitian ini, data yang divisualisasikan adalah DataFrame `gini_data`, dengan sumbu x mewakili variabel "Tahun" dan sumbu y mewakili variabel "Gini Rasio". Scatter plot ini memberikan gambaran awal tentang distribusi data dan hubungan antara tahun dan nilai Gini Rasio sebelum dilakukan pengelompokan atau analisis lebih lanjut. Dengan memvisualisasikan data dalam bentuk bagan koordinat, pengguna dapat dengan mudah mengidentifikasi pola, tren, dan kemungkinan anomali dalam data, yang merupakan langkah penting dalam eksplorasi data dan persiapan untuk analisis klastering.

4.2.5. Proses Input Centroid

```
# $$ Fungsi untuk menginput jumlah klaster yang akan digunakan
def get_number_input():
    while True:
        try:
            # Asking the user to enter a number
            number = int(input("Jumlah klaster yang akan diimplementasi: "))
            return number
        except ValueError:
            # Handling the case where the input is not a valid number
            print("------Yang anda masukkan bukanlah angka, mohon diulang------")

# Main function
user_number = get_number_input()
print(f"\n>>>Jumlah Klaster yang akan digunakan: {user_number} klaster")
print("Silahkan lanjutkan pada proses selanjutnya")
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.5. Kode Fitur Input Centroid

Kode tahapan ini terdiri dari sebuah fungsi yang digunakan untuk meminta pengguna memasukkan jumlah klaster yang akan digunakan dalam analisis klastering. Fungsi `get_number_input()` menjalankan loop yang terus berulang hingga pengguna memasukkan angka valid (integer).

Pada setiap iterasi, program meminta input dari pengguna dengan pesan "Jumlah klaster yang akan diimplementasi: ". Jika pengguna memasukkan nilai yang tidak valid (bukan angka), program akan menganggapnya sebagai `ValueError` dan menampilkan pesan pernyataan error yang berisi "------Yang anda masukkan bukanlah angka, mohon diulang------", kemudian meminta input kembali, ketika pengguna memasukkan angka yang valid, maka data inputannya akan ditampilkan kembali.

Selanjutnya, nilai yang dimasukkan pengguna disimpan dalam variabel `user_number`, dan program mencetak jumlah klaster yang akan digunakan dengan format: `>>>Jumlah Klaster yang akan digunakan: {user_number} klaster`. Setelah

itu, program memberikan instruksi kepada pengguna untuk melanjutkan ke proses berikutnya. Kode ini memastikan bahwa input yang diterima adalah valid dan sesuai dengan yang dibutuhkan untuk proses klastering.

4.2.6. Proses Penampilan Koordinat Centroid Iterasi Pertama

```
class CustomKMeans(KMeans):
    def _init_centroids(self, X, x_squared_norms, init, random_state, init_size=None):
        if init == 'random':
            n_samples = X.shape[0]
            seeds = random_state.permutation(n_samples)[:self.n_clusters]
            centers = X[seeds]
            self.initial_centroids_ = centers.copy()
            return centers
        else:
            return super()._init_centroids(X, x_squared_norms, init, random_state, init_size)

    def _assign_labels(self, X):
        self.labels_, _ = pairwise_distances_argmin_min(X, self.cluster_centers_, metric='manhattan')

# Iterasi Pertama dengan custom KMeans
kmeans = CustomKMeans(n_clusters=user_number, init='random', random_state=0, n_init='auto')
kmeans.fit(gini_data[['Tahun', 'Gini Rasio']])

# Menampilkan centroid awal yang diinisialisasi secara acak
initial_centroids = kmeans.initial_centroids_
print("Centroid iterasi pertama:")
print(initial_centroids)
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.6. Kode Fitur Penampilan Koordinat Centroid Iterasi Pertama

Kode tersebut merupakan implementasi dari sebuah kelas yang disebut `CustomKMeans`, yang merupakan turunan dari kelas `KMeans` yang ada di dalam library scikit-learn. Kelas ini memiliki dua metode tambahan yang dioverride dari kelas induknya.

Metode pertama, `_init_centroids`, digunakan untuk menginisialisasi posisi centroid. Jika parameter `init` yang diterima adalah 'random', maka centroid akan diambil secara acak dari sampel data yang tersedia. Jumlah centroid yang diambil sesuai dengan jumlah klaster yang ditentukan. Jika `init` tidak 'random', metode ini akan memanggil metode induknya dari kelas `KMeans`.

Metode kedua, `_assign_labels`, digunakan untuk menentukan label klaster untuk setiap titik data berdasarkan jarak Manhattan terdekat antara titik data tersebut dengan centroid klaster, dengan mendefinisikan kelas `CustomKMeans`, kode selanjutnya menggunakan kelas ini untuk melakukan iterasi pertama dari algoritma K-Means.

4.2.7. Proses Clustering

```
# Custom K-Means with Manhattan distance
class KMeansManhattan(KMeans):
    def __init__(self, n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=1e-4, random_state=None):
        super().__init__(n_clusters=n_clusters, init=init, n_init=n_init, max_iter=max_iter, tol=tol, random_state=random_state)

    def _e_step(self, X):
        labels = np.argmin(manhattan_distances(X, self.cluster_centers_), axis=1)
        return labels

# Membuat data gini_data (asumsi data sudah tersedia)
# gini_data = ...

# Menggunakan K-Means dengan Manhattan distance
kmeans = KMeansManhattan(n_clusters= user_number, random_state=0)
kmeans.fit(gini_data)

# Membuat plot 2D
fig, ax = plt.subplots()

# Plot data dengan warna berdasarkan label cluster
scatter = ax.scatter(gini_data['Tahun'], gini_data['Gini Rasio'], c=kmeans.labels_, cmap='viridis')

# Plot centroid yang telah dikluster dengan simbol bintang dan warna merah
for i, centroid in enumerate(kmeans.cluster_centers_):
    ax.scatter(centroid[0], centroid[1], color='red', marker='*', s=100, label=f'Centroid Klaster {i}')

# Labeling axes
ax.set_xlabel('Tahun')
ax.set_ylabel('Gini Rasio')
ax.set_title('K-Means Clustering Plot Manhattan')

# Menambahkan legenda
#ax.legend()

# Menampilkan plot
plt.show()

# Jumlah iterasi
iterations = kmeans.n_iter_
print(f"Jumlah iterasi yang terjadi pada K-Means: {iterations} iterasi")
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.7. Kode Proses Clustering

Kode pada proses ini berguna untuk mengimplementasikan algoritma K-Means dengan menggunakan metode Manhattan sebagai metrik *distance*. Langkah pertama adalah membuat objek KMeansManhattan dengan menentukan jumlah klaster yang diinginkan dan titik acak. Setelah kode bertemu dengan data yang diberikan, dilakukan visualisasi hasil klastering dalam plot 2D (tabel koordinat). Dalam plot tersebut, setiap titik data direpresentasikan dengan sumbu x dan sumbu

y, yang masing-masing mewakili variabel "Tahun" dan "Gini Rasio" dari dataset. Warna titik-titik data ditentukan berdasarkan klaster yang diberikan oleh algoritma K-Means. Selain itu, centroid klaster ditampilkan sebagai simbol bintang berwarna merah di plot, di mana posisi centroid ini telah dihitung dan disesuaikan dengan klaster yang terbentuk.

Selanjutnya, label sumbu dan judul plot ditambahkan untuk memberikan konteks pada visualisasi yang dihasilkan. Ada juga kemungkinan untuk menambahkan legenda, meskipun dalam kode ini peneliti membuat kode tersebut di-comment-kan (dinonaktifkan) karena menghalangi grafik pada plot. Plot ditampilkan kepada pengguna menggunakan perintah `plt.show()`, selain visualisasi, kode diatas juga mencetak jumlah iterasi yang terjadi selama proses klastering menggunakan metode K-Means. Informasi iterasi memberikan gambaran tentang kompleksitas algoritma dan seberapa cepat atau lambat konvergensi dilakukan untuk dataset yang diberikan. Dengan demikian, dengan menggunakan kode yang dibuat diatas maka pengguna dapat melakukan analisis klastering data dengan memanfaatkan metrik *distance* Manhattan dan memvisualisasikan hasilnya dalam plot 2D dengan memanfaatkan Python.

4.2.8. Proses Dunn Index

```python
# Menghitung Dunn Index
from sklearn.metrics.pairwise import pairwise_distances

def calculate_dunn_index(X, labels, centroids):
    # Hitung jarak antar titik di dalam setiap klaster
    intra_cluster_distances = []
    for cluster_label in set(labels):
        cluster_points = X[labels == cluster_label]
        intra_cluster_distances.append(np.max(pairwise_distances(cluster_points, metric='manhattan')))

    # Hitung jarak antara klaster
    inter_cluster_distances = pairwise_distances(centroids, metric='manhattan')

    # Hitung Dunn Index
    max_intra_cluster_distance = np.max(intra_cluster_distances)
    min_inter_cluster_distance = np.min(inter_cluster_distances[np.nonzero(inter_cluster_distances)])

    dunn_index = min_inter_cluster_distance / max_intra_cluster_distance
    return dunn_index

# Menghitung Dunn Index
dunn_index = calculate_dunn_index(gini_data[['Tahun', 'Gini Rasio']], kmeans.labels_, initial_centroids)
print(f"Dunn Index dari hasil klastering: {dunn_index}")
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.8. Kode Proses Dunn Index

Kode pada tahapan ini bertujuan untuk menghitung Dunn Index sebagai metrik evaluasi untuk hasil klastering yang telah dilakukan sebelumnya menggunakan algoritma K-Means dengan *distance* Manhattan. Dunn Index adalah sebuah metrik yang digunakan untuk mengevaluasi kualitas klastering dengan mempertimbangkan rasio antara jarak terpendek antara klaster dengan jarak terjauh di dalam klaster. Prosedur perhitungan Dunn Index dimulai dengan menghitung jarak antar titik di dalam setiap klaster. Ini dilakukan dengan mencari jarak terjauh antara dua titik di dalam klaster menggunakan metrik *distance* Manhattan. Kemudian, jarak antara klaster dihitung dengan menghitung jarak antara centroid masing-masing klaster menggunakan metrik *distance* Manhattan.

Proses berlanjut setelah mendapatkan jarak antar centroid klaster dan jarak antar data dalam masing-masing klaster, Dunn Index dihitung dengan membagi jarak terpendek antar klaster dengan jarak terjauh di dalam masing-masing klaster. Semakin besar nilai Dunn Index, semakin baik kualitas klasteringnya. Fungsi

`calculate_dunn_index` digunakan untuk menjalankan proses perhitungan Dunn Index.

Setelah Dunn Index dihitung, hasilnya dicetak untuk memberikan informasi tentang kualitas klastering yang telah dilakukan. Ini memberikan pemahaman tambahan tentang seberapa baik klastering telah berhasil memisahkan data ke dalam kelompok-kelompok yang terpisah. Dengan demikian, kode ini memberikan alat evaluasi yang berguna untuk mengevaluasi kualitas klastering yang dilakukan dengan menggunakan algoritma K-Means dengan metode jarak / *distance* Manhattan.

4.3. Pengujian Sistem Data Mining K-means Clustering Menggunakan Euclidean Distance

4.3.1. Proses Input Library

```
#Import library Pandas untuk mengolah file yang akan dianalisis
import pandas as pd
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples, silhouette_score
import matplotlib.pyplot as plt
from sklearn.metrics import silhouette_score
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.9. Kode Input Library

Proses clustering dimulai dengan memanggil library sebagai fungsi untuk menjalankan proses analisis pada file yang akan digunakan, adapun library yang digunakan ialah library Pandas dan Seaborn dengan cara mengimpornya. Penggunaan sklearn.cluster dan sklearn.metrics dimanfaatkan agar proses algoritma K-means dapat dilakukan. Impor berikutnya ialah matplotlib.pyplot yang

mana berguna untuk menampilkan grafik koordinat yang akan digunakan sebagai wadah visualisasi clustering. Impor terakhir yakni ialah numpy yang mana tujuannya agar pengguna dapat menggunakan sklearn.metrics.pairwise yang bermanfaat untuk melakukan proses *Euclidean Distance.*

4.3.2. Proses Input Data File

```
# Mengupload file CSV untuk diolah (gini_rasio_a.csv)
from google.colab import files
uploaded = files.upload()
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.10. Kode Fitur Input Data

Kode pada tahapan ini digunakan untuk mengunggah file CSV bernama "gini_rasio_a.csv" dari komputer pengguna ke program yang sedang dibangun. Fungsi `files.upload()` dari modul `google.colab` memungkinkan pengguna untuk memilih file dari komputer pengguna dan mengunggahnya ke sistem *Colab* yang sedang berjalan. Ini adalah langkah awal yang sering digunakan dalam proses analisis data untuk mengimpor dataset yang akan dianalisis. Setelah file diunggah, dataset akan tersedia dalam bentuk objek file yang dapat diakses dan dimanipulasi menggunakan library seperti Pandas. Dengan mengunggah file CSV ini, pengguna dapat melakukan berbagai operasi analisis data seperti pemrosesan data mining dan visualisasinya. Proses ini mempermudah pengguna untuk membawa data dari komputer pengguna ke dalam sistem cloud *Google Colab.*

4.3.3. Proses Pengecekan Data

```
#Mengecek isian file yang akan diolah
gini_data = pd.read_csv('gini_rasio_a.csv', delimiter=';', usecols = ['Tahun', 'Gini Rasio'])
gini_data.head()
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.11. Kode Fitur Pengecekan Data

Kode di atas digunakan untuk membaca dan menampilkan isi awal dari file CSV bernama "gini_rasio_a.csv" yang telah diunggah sebelumnya. Library Pandas menyediakan fungsi `read_csv` untuk membaca file CSV, dan dalam hal ini, file dibaca dengan pemisah delimiter titik koma (`;`). Parameter `usecols` digunakan untuk memilih hanya kolom tertentu, yaitu "Tahun" dan "Gini Rasio", yang relevan untuk analisis yang akan dilakukan. Setelah data dibaca ke dalam DataFrame Pandas yang dinamai `gini_data`, fungsi `head()` dipanggil untuk menampilkan lima baris pertama dari DataFrame tersebut. Ini memungkinkan pengguna untuk memverifikasi bahwa data telah diimpor dengan benar dan untuk melihat sekilas struktur dan konten data yang akan dianalisis. Langkah ini penting untuk memastikan bahwa data sesuai dengan ekspektasi sebelum melakukan analisis lebih lanjut.

4.3.4. Proses Pengecekan Data dengan tabel koordinat

```
#Memunculkan data yang akan diolah kedalam bentuk bagan koordinat, bagan pra-pengelompokkan.
sns.scatterplot(data = gini_data, x = 'Tahun', y = 'Gini Rasio')
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.12. Kode Fitur Pengecekan Data Dengan Tabel Koordinat

Kode di atas digunakan untuk membuat visualisasi data dalam bentuk tabel koordinat scatter plot menggunakan library Seaborn. Fungsi `scatterplot` dari Seaborn membantu memvisualisasikan hubungan antara dua variabel dalam dataset. Dalam penelitian ini, data yang divisualisasikan adalah DataFrame `gini_data`, dengan sumbu x mewakili variabel "Tahun" dan sumbu y mewakili variabel "Gini Rasio". Scatter plot ini memberikan gambaran awal tentang distribusi data dan hubungan antara tahun dan nilai Gini Rasio sebelum dilakukan pengelompokan atau analisis lebih lanjut. Dengan memvisualisasikan data dalam

bentuk bagan koordinat, pengguna dapat dengan mudah mengidentifikasi pola, tren, dan kemungkinan anomali dalam data, yang merupakan langkah penting dalam eksplorasi data dan persiapan untuk analisis klastering.

4.3.5. Proses Input Centroid

```
# $$ Fungsi untuk menginput jumlah klaster yang akan digunakan
def get_number_input():
    while True:
        try:
            # Asking the user to enter a number
            number = int(input("Jumlah klaster yang akan diimplementasi: "))
            return number
        except ValueError:
            # Handling the case where the input is not a valid number
            print("------Yang anda masukkan bukanlah angka, mohon diulang------")

# Main function
user_number = get_number_input()
print(f"\n>>>Jumlah Klaster yang akan digunakan: {user_number} klaster")
print("Silahkan lanjutkan pada proses selanjutnya")
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.13. Kode Fitur Input Centroid

Kode tahapan ini terdiri dari sebuah fungsi yang digunakan untuk meminta pengguna memasukkan jumlah klaster yang akan digunakan dalam analisis klastering. Fungsi `get_number_input()` menjalankan loop yang terus berulang hingga pengguna memasukkan angka valid (integer).

Pada setiap iterasi, program meminta input dari pengguna dengan pesan "Jumlah klaster yang akan diimplementasi: ". Jika pengguna memasukkan nilai yang tidak valid (bukan angka), program akan menganggapnya sebagai `ValueError` dan menampilkan pesan pernyataan error yang berisi "------Yang anda masukkan bukanlah angka, mohon diulang------", kemudian meminta input kembali, ketika pengguna memasukkan angka yang valid, maka data inputannya akan ditampilkan kembali.

Selanjutnya, nilai yang dimasukkan pengguna disimpan dalam variabel `user_number`, dan program mencetak jumlah klaster yang akan digunakan dengan format: `>>>Jumlah Klaster yang akan digunakan: {user_number} klaster`. Setelah itu, program memberikan instruksi kepada pengguna untuk melanjutkan ke proses berikutnya. Kode ini memastikan bahwa input yang diterima adalah valid dan sesuai dengan yang dibutuhkan untuk proses klastering.

4.3.6. Proses Penampilan Koordinat Centroid Iterasi Pertama

```
#Pengolahan K-Means, menggunakan EUCLIDEAN DISTANCE (Default sklearn.cluster)
class CustomKMeans(KMeans):
    def _init_centroids(self, X, x_squared_norms, init, random_state, init_size=None):
        if init == 'random':
            n_samples = X.shape[0]
            seeds = random_state.permutation(n_samples)[:self.n_clusters]
            centers = X[seeds]
            self.initial_centroids_ = centers.copy()
            return centers
        else:
            return super()._init_centroids(X, x_squared_norms, init, random_state, init_size)

# Iterasi Pertama dengan custom KMeans
kmeans = CustomKMeans(n_clusters=user_number, init='random', random_state=0, n_init='auto')
kmeans.fit(gini_data[['Tahun', 'Gini Rasio',]])

# Menampilkan centroid awal yang diinisialisasi secara acak
initial_centroids = kmeans.initial_centroids_
print("Centroid iterasi pertama:")
print(initial_centroids)
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.14. Kode Penampilan Koordinat Centroid Iterasi Pertama

Kode di atas memiliki fungsi untuk memperlihatkan implementasi dari algoritma K-Means dengan menggunakan jarak Euclidean distance sebagai metrik *distance* bawaan pada iterasi pertama. Kode ini menggunakan pendekatan yang disesuaikan dengan mendefinisikan sebuah kelas baru yang disebut `CustomKMeans`, yang merupakan turunan dari kelas KMeans yang ada di dalam library scikit-learn. Pada kelas ini, dilakukan override terhadap metode `_init_centroids`, yang bertanggung jawab untuk menginisialisasi posisi centroid.

Dalam metode ini, jika parameter `init` yang diberikan adalah 'random', maka posisi centroid akan dipilih secara acak dari sampel data yang ada. Setelah model K-Means dibuat dengan menggunakan kelas `CustomKMeans` dan dilatih dengan data yang diberikan, posisi centroid awal yang diinisialisasi secara acak akan ditampilkan ke layar sebagai informasi tambahan untuk proses klastering.

4.3.7. Proses Clustering

```
# Membuat plot 2D
fig, ax = plt.subplots()

# Plot data dengan warna berdasarkan label cluster
scatter = ax.scatter(gini_data['Tahun'], gini_data['Gini Rasio'], c=kmeans.labels_, cmap='viridis')

# Plot centroid yang telah dikluster dengan simbol bintang dan warna merah
for i, centroid in enumerate(kmeans.cluster_centers_):
    ax.scatter(centroid[0], centroid[1], color='red', marker='*', s=100, label=f'Centroid Klaster {i}')

# Labeling axes
ax.set_xlabel('Tahun')
ax.set_ylabel('Gini Rasio')
ax.set_title('K-Means Clustering Plot Euclidean')

# Menambahkan legenda
#ax.legend()

# Menampilkan plot
plt.show()

# Jumlah iterasi
iterations = kmeans.n_iter_
print(f"Jumlah iterasi yang terjadi pada K-Means: {iterations} iterasi")
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.15. Kode Proses Clustering

Kode diatas memiliki tujuan untuk membuat visualisasi plot 2D dari hasil klastering yang telah dilakukan menggunakan algoritma K-Means dengan jarak Euclidean distance. Pertama, plot 2D dibuat dengan menggunakan `plt.subplots()`. Data yang telah dikelompokkan ke dalam klaster direpresentasikan dalam plot dengan warna yang berbeda-beda untuk setiap klaster, menggunakan label klaster yang diberikan oleh algoritma K-Means. Posisi centroid klaster kemudian juga ditampilkan di dalam plot sebagai simbol bintang berwarna merah. Proses ini

dilakukan dengan iterasi melalui setiap centroid yang telah dihitung oleh algoritma K-Means, kemudian menambahkannya ke dalam plot.

Setelah menampilkan data dan centroid di plot, sumbu x dan y diberi label sesuai dengan variabel yang mewakili data, yaitu "Tahun" dan "Gini Rasio". Plot ini juga diberi judul yang menjelaskan konteks visualisasi yang dilakukan, yaitu "K-Means Clustering Plot Euclidean". Selanjutnya, jumlah iterasi yang terjadi selama proses klastering juga dicetak ke layar sebagai informasi tambahan. Dengan menggunakan kode ini, pengguna dapat dengan mudah memvisualisasikan hasil klastering yang telah dilakukan dan mendapatkan informasi tambahan tentang proses iterasi yang terjadi dalam algoritma K-Means yang menggunakan Euclidean *distance* dengan melihat pada berapa jumlah iterasi yang terjadi.

4.3.8. Proses Dunn Index

```python
#Menghitung Dunn Index
from sklearn.metrics.pairwise import euclidean_distances

def calculate_dunn_index(X, labels, centroids):
    # Hitung jarak antar titik di dalam setiap klaster
    intra_cluster_distances = []
    for cluster_label in set(labels):
        cluster_points = X[labels == cluster_label]
        intra_cluster_distances.append(np.max(euclidean_distances(cluster_points, cluster_points)))

    # Hitung jarak antara klaster
    inter_cluster_distances = euclidean_distances(centroids, centroids)

    # Hitung Dunn Index
    max_intra_cluster_distance = np.max(intra_cluster_distances)
    min_inter_cluster_distance = np.min(inter_cluster_distances[np.nonzero(inter_cluster_distances)])

    dunn_index = min_inter_cluster_distance / max_intra_cluster_distance
    return dunn_index

# Menghitung Dunn Index
dunn_index = calculate_dunn_index(gini_data[['Tahun', 'Gini Rasio']], kmeans.labels_, initial_centroids)
print(f"Dunn Index dari hasil klastering: {dunn_index}")
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.16. Kode Dunn Index

Kode di atas memiliki fungsi untuk menghitung Dunn Index, yang merupakan metrik evaluasi untuk mengevaluasi kualitas klastering yang telah dilakukan.

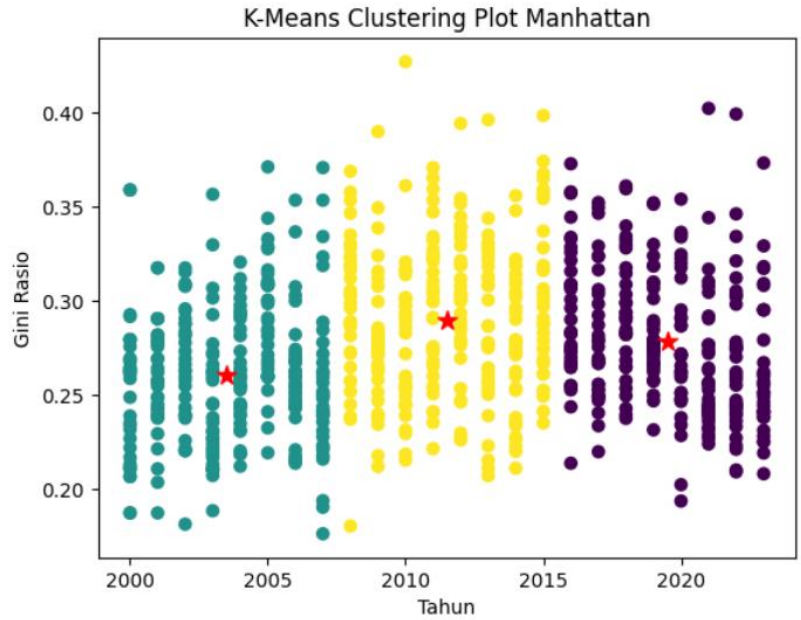
Pertama, fungsi `calculate_dunn_index` digunakan untuk menjalankan proses kalkulasi untuk mendapatkan nilai Dunn Index yang ingin dilihat. Metode ini menghitung jarak intra-klaster dengan mengidentifikasi jarak terjauh antara dua titik dalam setiap klaster menggunakan jarak Euclidean. Kemudian, jarak antar-klaster dihitung dengan mencari jarak antara setiap pasangan centroid klaster. Setelah mendapatkan jarak intra-klaster dan jarak antar-klaster, Dunn Index dihitung dengan membagi jarak terpendek antar klaster dengan jarak terjauh di dalam klaster. Nilai Dunn Index yang lebih besar menunjukkan kualitas klastering yang lebih baik.

Setelah Dunn Index dihitung, hasilnya dicetak ke layar sebagai informasi tambahan tentang kualitas klastering yang telah dilakukan. Hal ini memberikan pemahaman tambahan tentang seberapa baik klastering telah berhasil memisahkan data ke dalam kelompok-kelompok yang terpisah. Dengan menggunakan metrik evaluasi seperti Dunn Index, pengguna dapat memperoleh wawasan yang lebih baik tentang performa algoritma klastering yang digunakan.

4.4. Pengujian Menggunakan Data Rasio Gini

Pada penelitian ini, peneliti akan melakukan pengujian kode clustering K-means yang telah dibuat dengan metode jarak Manhattan dan Euclidean. Kedua metode tersebut akan diuji menggunakan data rasio Gini dari seluruh Kabupaten / Kota di Sumatera Utara mulai dari tahun 2000 sampai dengan tahun 2023 guna melihat bagaimana pengelompokkannya dan melihat performa kedua metode dari aspek Dunn Index dan jumlah iterasi agar dapat mengetahui metode mana yang paling efektif untuk mengelolah data Dunn Index dengan catatan cluster yang akan dibentuk ada sebanyak 3 cluster.

4.4.1. Pengujian K-Means Clustering dengan metode Manhattan



Sumber: Dokumentasi Peneliti, 2024

Gambar 4.17. Hasil Clustering Menggunakan Metode Manhattan Distance

Berdasarkan data clustering yang diperoleh setelah melalui beberapa tahap, dapat kita peroleh bahwasannya iterasi yang terjadi ada sebanyak 3 kali iterasi dimana setiap centroid cukup berjauhan satu sama lain.

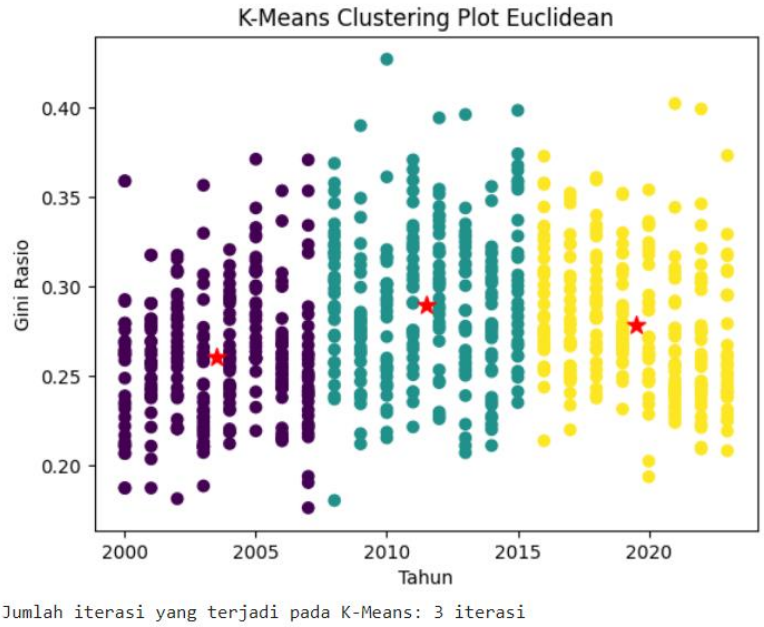
```
Dunn Index dari hasil klastering: 0.2860941592407533
```

Sumber: Dokumentasi Peneliti, 2024

Gambar 4.18. Hasil Dunn Index Menggunakan Metode Manhattan Distance

Adapun hasil perhitungannya diperoleh nilai sebesar 0.2860941592407533 dengan menggunakan kode yang telah dirancang sebelumnya setelah proses clustering berhasil dilakukan.

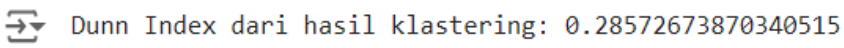
4.4.2. Pengujian K-Means Clustering dengan metode Euclidean



Sumber: Dokumentasi Peneliti, 2024

Gambar 4.19. Hasil Clustering Menggunakan Metode Euclidean Distance

Berdasarkan data clustering yang diperoleh setelah melalui beberapa tahap, dapat kita peroleh bahwasannya iterasi yang terjadi ada sebanyak 3 kali iterasi dimana setiap centroid cukup berjauhan satu sama lain, data yang diperoleh lumayan menyerupai hasil yang dikeluarkan oleh metode Manhattan namun pada Dunn Index terdapat perbedaan.



Sumber: Dokumentasi Peneliti, 2024

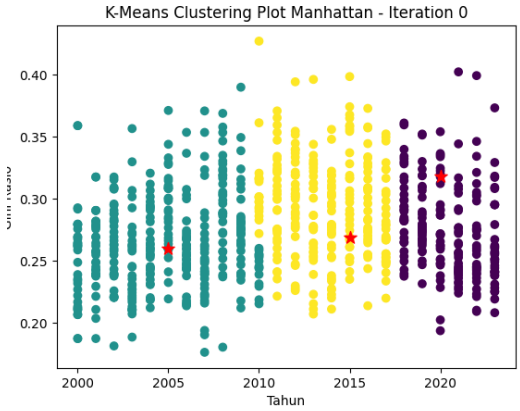
Gambar 4.20. Hasil Dunn Index Menggunakan Metode Euclidean Distance

Adapun hasil perhitungannya diperoleh nilai sebesar 0.28572673870340515 dengan menggunakan kode yang telah dirancang sebelumnya setelah proses clustering berhasil dilakukan. Nilai Dunn Index yang dihasilkan melalui penerapan

metode Euclidean tidak begitu jauh berbeda dari metode Manhattan namun posisinya sedikit lebih rendah dari Manhattan.

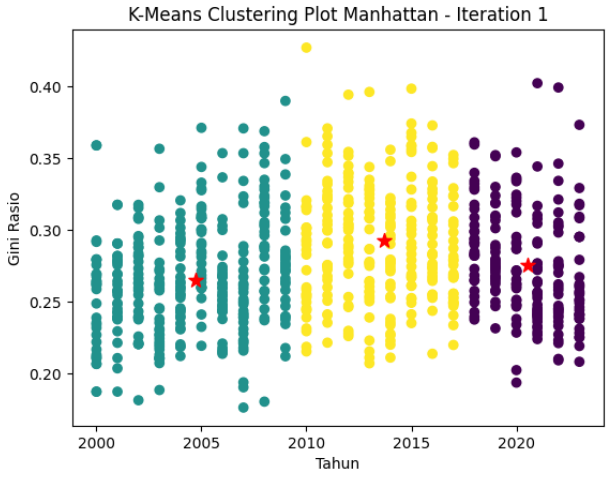
4.5. Interprestasi

Penelitian ini menghasilkan interprestasi yang dapat memberikan bagaimana penyebaran centroid pada setiap iterasi di kedua metode. Adapun grafiknya yakni sebagai berikut.



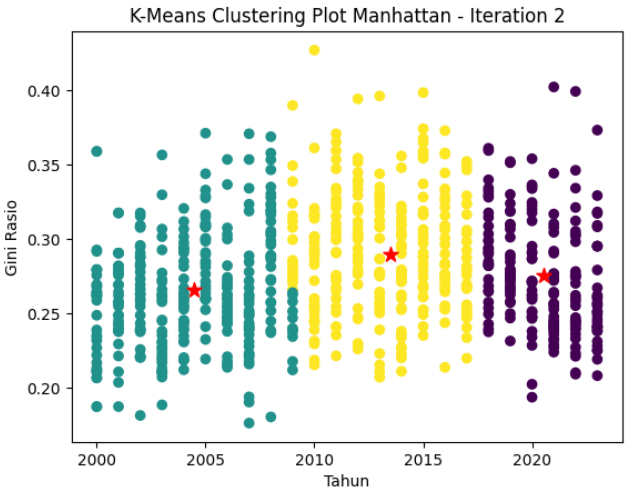
Sumber: Dokumentasi Peneliti, 2024

Gambar 4.21. Iterasi 0 Manhattan Distance



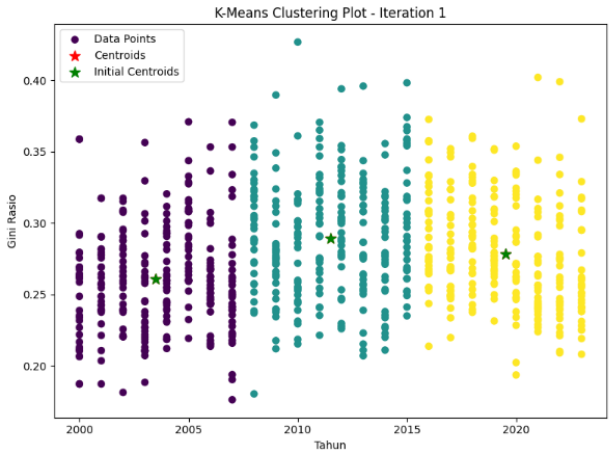
Sumber: Dokumentasi Peneliti, 2024

Gambar 4.22. Iterasi 1 Manhattan Distance



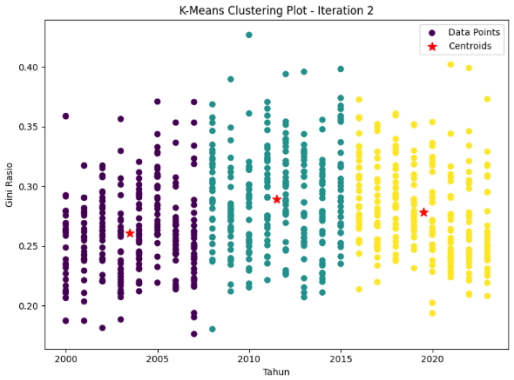
Sumber: Dokumentasi Peneliti, 2024

Gambar 4.23. Iterasi 2 Manhattan Distance



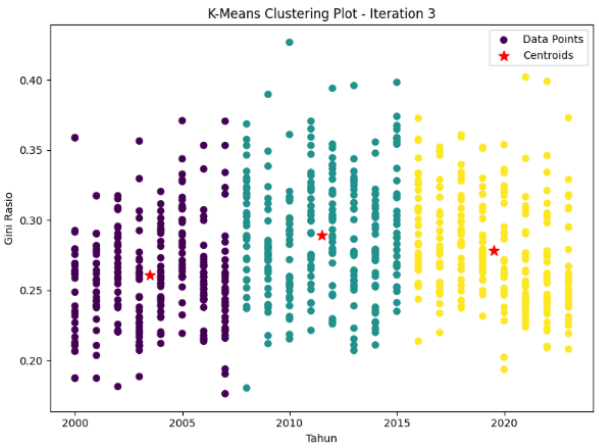
Sumber: Dokumentasi Peneliti, 2024

Gambar 4.24. Iterasi 0 Euclidean Distance



Sumber: Dokumentasi Peneliti, 2024

Gambar 4.25. Iterasi 1 Euclidean Distance



Sumber: Dokumentasi Peneliti, 2024

Gambar 4.26. Iterasi 2 Euclidean Distance

Berdasarkan hasil yang diperoleh, dapat dipahami bahwa penyebaran titik centroid dengan menggunakan Manhattan Distance cenderung lebih menyebar dan berbeda di setiap iterasi. Sebaliknya, pada Euclidean Distance, titik-titik centroid cenderung hampir sangat menyerupai satu sama lain di setiap iterasi apabila dilakukan secara otomatis menggunakan program. Hal ini mungkin diakibatkan oleh perbedaan fundamental antara kedua metode pengukuran jarak tersebut. Manhattan Distance mengukur jarak antara dua titik dengan mengikuti sumbu koordinat secara horizontal dan vertikal, yang menghasilkan jalur berbentuk grid.

Hal ini menyebabkan variasi yang lebih besar dalam distribusi centroid karena perbedaan kecil dalam data dapat menghasilkan perubahan yang lebih signifikan pada hasil akhirnya. Di sisi lain, Euclidean Distance mengukur jarak lurus antara dua titik, yang cenderung menghasilkan hasil yang lebih seragam dan stabil, karena perhitungan ini lebih sensitif terhadap jarak terpendek langsung antar titik. Oleh karena itu, dalam konteks iterasi otomatis, penggunaan Euclidean Distance memberikan hasil yang lebih konsisten dan seragam, sementara Manhattan Distance cenderung menghasilkan variasi yang lebih beragam pada setiap iterasi. Perbedaan ini penting untuk dipertimbangkan dalam pemilihan metode pengukuran jarak yang sesuai berdasarkan tujuan dan karakteristik dataset yang digunakan dalam penelitian.

BAB V

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil analisis K-means Clustering menggunakan Manhattan Distance dan Euclidean Distance terhadap data rasio Gini Sumatera Utara dari tahun 2000-2023 dapat diperoleh kesimpulan sebagai berikut:

1. Perbandingan kinerja Metode Manhattan dan Metode Euclidean dalam analisis data rasio Gini Sumatera Utara menggunakan algoritma K-means memilik kinerja yang kurang lebih menyerupai satu sama lain dalam hal pemprosesan pada kode namun terdapat perbedaan pada kode pengaktifannya tetapi pada kebutuhan library cendrung menyerupai satu sama lain.
2. Tidak terdapat perbedaan signifikan antara hasil distance yang diperoleh dari metode Manhattan dan Euclidean terkait pemahaman pola rasio Gini di Sumatera Utara.
3. Metode yang lebih unggul dalam mengukur distance ketika menghitung jarak data rasio Gini kabupaten / kota di Sumatera Utara ialah Manhattan Distance, meskipun jumlah iterasinya sama dengan Euclidean Distance yakni sebesar 3 iterasi namun dari Dunn Index Manhattan distance cendrung lebih unggul dengan nilai sebesar 0.2860941592407533 dan Euclidean distance sedikit lebih rendah sebesar 0.28572673870340515.

5.2. Saran

Saran dari penulis untuk peneliti di masa yang akan datang apabila berkecimpung dalam penelitian yang sejenis adalah:

1. Peneliti berikutnya dapat menggunakan algoritma yang lebih banyak sehingga akan terdapat lebih banyak pula perbandingan dimana bisa terdiri dari lebih dari satu algoritma berbeda dengan masing-masing 2 atau lebih pengujian distance agar memperoleh hasil temuan yang lebih banyak dan baru.
2. Peneliti berikutnya dapat menggunakan data yang lebih banyak seperti data rasio gini di seluruh kawasan Indonesia agar memperoleh temuan hasil performa distance yang lebih baik dalam gambaran yang lebih luas.
3. Peneliti berikutnya dapat mengimplementasikan proses clustering dan pengujian distance menggunakan perangkat lunak pembuatan aplikasi berbasis objek agar pengguna yang kurang paham penggunaan secara teknis google colab dapat lebih nyaman dalam melakukan pengujian.

DAFTAR PUSTAKA

Agus, Z. (2023). Sistem Pakar Diagnosa Kecanduan Game Online dengan Metode Certainty Factor dan Euclidean Distance Berbasis Web. *JDMIS: Journal of Data Mining and Information System*, *1*(1), 29–36. https://doi.org/10.54259/jdmis.v1i1.1521

Andika Naufal, M., Setiadi, B., & Trisnawijayana. (2022). Perhitungan Jarak Dalam Sistem Deteksi Social Distancing Dengan Menggunakan Metode Euclidean Distance. *Prosiding The 13th Industrial Research Workshop and National Seminar*, *13*(1), 1431–1435.

Anjumi, K. N., Bella, C., Komputer, T., Merapi, J., Mataram, F., Mataram, K. S., & Tengah, K. L. (2022). *Analisis Data Pola Penjualan Menggunakan*. *2*(2), 1–9.

Dinata, R. K., Akbar, H., & Hasdyna, N. (2020). Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus. *ILKOM Jurnal Ilmiah*, *12*(2), 104–111. https://doi.org/10.33096/ilkom.v12i2.539.104-111

Djouzi, K., Beghdad-bey, K., & Amamra, A. (2023). Big Data Sampling Techniques : A State-of-the-art Survey Big Data Sampling Techniques : A State-of-the-art Survey. *ResearchGate*, (June), 1–11.

El Islami, M. F., & Fitrianto, A. R. (2023). Pengaruh Penyaluran Dana ZIS, Inflasi, Dan Gini Ratio Terhadap Tingkat Kedalaman Kemiskinan Satu Dekade. *Jurnal Ilmiah Ekonomi Islam*, *9*(1), 229–239. Retrieved from http://dx.doi.org/10.29040/jiei.v9i1.6994

Eviana, A., Fauzan, A. C., Harliana, H., & Putra, F. N. (2022). Komparasi Jarak Euclidean dan Jarak Manhattan Untuk Deteksi Covid-19 Melalui Citra CT-Scan Paru-Paru. *Komputika : Jurnal Sistem Komputer*, *11*(2), 121–129. https://doi.org/10.34010/komputika.v11i2.5380

Ghazal, T. M., Hussain, M. Z., Said, R. A., Nadeem, A., Hasan, M. K., Ahmad, M., … Naseem, M. T. (2021). Performances of k-means clustering algorithm with different distance metrics. *Intelligent Automation and Soft Computing*, *30*(2), 735–742. https://doi.org/10.32604/iasc.2021.019067

H, N. A., Wijaya, K., Rahmanti, N., Kurnia, R., Ulyani, R., & ... (2023). Implementasi Algoritma Naïve Bayes untuk Memprediksi Penjualan Lampu Pada Toko Satria. *Innovative: Journal Of …*, *3*(2), 9373–9387. Retrieved from http://j-innovative.org/index.php/Innovative/article/view/1330

Hamdani, R., & Mayshelly, E. (2023). Kinerja perekonomian daerah dan kesejahteraan masyarakat di Daerah Istimewa Yogyakarta Tahun 2016-2020. *Proceeding of National Conference on Accounting & Finance*, *5*(2019), 10–25. https://doi.org/10.20885/ncaf.vol5.art2

Hartono, B., Eniyati, S., & Hadiono, K. (2023). Perbandingan Metode Perhitungan Jarak pada Nilai Centroid dan Pengelompokan Data Menggunakan K-Means Clustering. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(3), 503–509. https://doi.org/10.30865/json.v4i3.6021

Hasugian, M., & Wardarita, R. (2022). Analisis Kajian Psikologi pada Film Tanah Surga Katanya. *Journal on Teacher Education*, *4*(2), 831–839.

Kurniawan, R., Hasibuan, M. S., & Hasibuan, R. (2023). Klasterisasi Wilayah Prioritas Vaksin Menggunakan Algoritma K-Means Clustering. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(3), 1585–1592. https://doi.org/10.30865/klik.v4i3.1334

Lani, O. P., Hendra, T., & Khalid, I. (2023). Peran Humas Badan Pusat Statistik (BPS) Provinsi Sumatera Barat dalam Mensosialisasikan Kegiatan Kegiatan Sensus Ekonomi 2016. *At-Tadabbur: Jurnal Penelitian Sosial Keagamaan*, *13*(1), 30–46.

Maknun, L., Syukur, A., Affandy, A., & Soeleman, M. A. (2022). Deteksi Dini Covid-19 Melalui Citra CT-Scan Paru-Paru Menggunakan K-Nearest Neighbor dengan Komparasi Jarak. *Jurnal Indonesia Sosial Teknologi*, *3*(3), 461–467. https://doi.org/10.36418/jist.v3i3.397

Martiano, M., Sari, Y., & Akbar, F. (2023). Analysis and Optimization of the K-Means Algorithm in Determining Course Scheduling. *Journal of Information System Research (JOSH)*, *5*(1), 134–141. https://doi.org/10.47065/josh.v5i1.4343

Mirantika, N., Syamfithriani, T. S., & Trisudarmo, R. (2023). Implementasi Algoritma K-Medoids Clustering Untuk Menentukan Segmentasi Pelanggan. *Jurnal Nuansa Informatika*, *17*(1), 196–204. Retrieved from https://journal.uniku.ac.id/index.php/ilkom

Mughnyanti, M., & Hafiz Nanda Ginting, S. (2023). Data Mining Manhattan Distance dan Euclidean Distance Pada Algoritma X-Means Dalam Klasifikasi Minat dan Bakat Siswa. *Remik*, *7*(1), 835–842. https://doi.org/10.33395/remik.v7i1.12162

Pangestu, M. S., & Fitriani, M. A. (2022). Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means. *Sainteks*, *19*(2), 141. https://doi.org/10.30595/sainteks.v19i2.14495

Rahayu, A. E., Fauzan, A. C., & Harliana, H. (2022). Komparasi Jarak Euclidean dan Manhattan Pada Algoritma K-Nearest Neighbor Dalam Mendeteksi Penyakit Diabetes Mellitus. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(2), 413–419. https://doi.org/10.30865/json.v4i2.5046

Septianingsih, A. (2022). Pemetaan Kabupaten Kota Di Provinsi Jawa Timur Berdasarkan Tingkat Kasus Penyakit Menggunakan Pendekatan Agglomeratif Hierarchical Clustering. *Jurnal Lebesgue : Jurnal Ilmiah Pendidikan Matematika, Matematika Dan Statistika*, *3*(2), 367–386.

https://doi.org/10.46306/lb.v3i2.139

Siahaan, M. (2022). Data Mining Strategi Pembangunan Infrastruktur Menggunakan Algoritma K-Means. *Jurnal Sisfokom (Sistem Informasi Dan Komputer)*, *11*(3), 316–324. https://doi.org/10.32736/sisfokom.v11i3.1453

Siregar, A. C., & Octariadi, B. C. (2021). Perbandingan Metode Jaringan Syaraf Tiruan Pada Klasifikasi Motif Kain Tenun Sambas. *Cybernetics*, *4*(02), 109–120. https://doi.org/10.29406/cbn.v4i02.2489

Solikhun, P. L. (2022). Comparison of Euclidean with Manhattan in K-Means Clustering for Grouping Palm Oil Production in the Province North Sumatra. *IJISTECH (International Journal of Information System and Technology)*, *5*(158), 709–716.

Sriadhi, S., Gultom, S., & Martiano, M. (2020). Accuracy of data cluster using modify k-mean algorithm by local deviation method. *International Journal of Advanced Science and Technology*, *29*(5), 2019–2025.

Suraya, S., Sholeh, M., & Andayati, D. (2023). Comparison of distance metric in k-mean algorithm for clustering wheat grain datasheet. *Jurnal Teknik Informatika C.I.T Medicom*, *15*(2), 73–83. https://doi.org/10.35335/cit.vol15.2023.408.pp73-83

Wijaya, D., & Widiarti, A. R. (2024). Batik classification using KNN algorithm and GLCM features extraction. *E3S Web of Conferences*, *475*, 1–13. https://doi.org/10.1051/e3sconf/202447502012

Yudha, D., Widodo, P., & Febiharsa, D. (2023). ANALISIS NILAI GIZI BALITA DI DESA MANGUNSARI KECAMATAN GUNUNGPATI KOTA SEMARANG DENGAN ALGORITMA K-MEANS CLUSTERING UNTUK PENCEGAHAN STUNTING Tumbuh kembang balita yang baik merupakan salah satu faktor pendukung kemajuan suatu negara . Akan tetapi persma. *EDUSTEMS*, *1*(1), 415–428.

Zai, C. (2022). Implementasi Data Mining Sebagai Pengolahan Data. *Portal Data*, *2*(3), 1–12. Retrieved from http://portaldata.org/index.php/portaldata/article/view/107

…

..

.

.

LAMPIRAN

Pengujian Manual Euclidean Distance Iterasi 0

Tahun	Gini Rasio	Cent 1X	Cent 1Y	Cent2X	Cent 2Y	Cent 3X	Cent 3Y
0	0,21661	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
0	0,21059	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
0	0,21284	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
0	0,26113	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,2755	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,21059	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,18719	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,26414	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,26177	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,27939	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060

Dist,C1	Dist, C2	Dist, C3	Cluster V.	MATCH	N_C1X	N_C1Y	N_C2X	N_C2Y	N_C3X	N_C3Y
1,467865	3,46111	5,45155	1,46786474	1	0,5	0,22153	0	0	0	0
1,466866	3,46057	5,4513	1,4668655	1						
1,467236	3,46077	5,45139	1,46723604	1						
1,475997	3,46541	5,45355	1,47599651	1						
2,465997	4,46246	6,45314	2,46599675	1						
2,458799	4,45753	6,45064	2,45879938	1						
2,45662	4,45598	6,44989	2,45661992	1						
2,464615	4,46153	6,45266	2,46461481	1						
2,464334	4,46134	6,45256	2,46433361	1						
2,466481	4,46278	6,45331	2,46648131	1						

Pengujian Manual Euclidean Distance Iterasi 1

Tahun	Gini Rasio	Cent 1X	Cent 1Y	Cent2X	Cent 2Y	Cent 3X	Cent 3Y
0	0,216613	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
0	0,210594	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
0	0,212843	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
0	0,26113	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
1	0,275501	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
1	0,210594	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
1	0,187192	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
1	0,264138	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
1	0,261772	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000
1	0,279388	0,5000	0,2215	0,0000	0,0000	0,0000	0,0000

Dist,C1	Dist, C2	Dist, C3	Cluster V.	MATCH	N_C1X	N_C1Y	N_C2X	N_C2Y	N_C3X	N_C3Y
0,500024	0,216613	0,216613	0,216613	2	1	0,246431	0	0,225295	0	0
0,50012	0,210594	0,210594	0,210594	2						
0,500076	0,212843	0,212843	0,212843	2						
0,501566	0,26113	0,26113	0,26113	2						
0,502904	1,037256	1,037256	0,502904	1						
0,50012	1,021934	1,021934	0,50012	1						
0,501178	1,01737	1,01737	0,501178	1						
0,501812	1,034296	1,034296	0,501812	1						
0,501617	1,033695	1,033695	0,501617	1						
0,503336	1,038296	1,038296	0,503336	1						

Pengujian Manual Euclidean Distance Iterasi 2

Tahun	Gini Rasio	Cent 1X	Cent 1Y	Cent2X	Cent 2Y	Cent 3X	Cent 3Y
0	0,216613	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
0	0,210594	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
0	0,212843	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
0	0,26113	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
1	0,275501	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
1	0,210594	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
1	0,187192	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
1	0,264138	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
1	0,261772	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
1	0,279388	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000

Dist,C1	Dist, C2	Dist, C3	Cluster V	MATCH	N_C1X	N_C1Y	N_C2X	N_C2Y	N_C3X	N_C3Y
1,000444	0,008682	0,216613	0,008682	2	1	0,246431	0	0,225295	0	0
1,000642	0,014701	0,210594	0,014701	2						
1,000564	0,012452	0,212843	0,012452	2						
1,000108	0,035835	0,26113	0,035835	2						
0,02907	1,00126	1,037256	0,02907	1						
0,035837	1,000108	1,021934	0,035837	1						
0,059239	1,000726	1,01737	0,059239	1						
0,017707	1,000754	1,034296	0,017707	1						
0,015341	1,000665	1,033695	0,015341	1						
0,032957	1,001462	1,038296	0,032957	1						

Pengujian Manual Manhattan Distance Iterasi 0

Tahun	Gini Rasio	Cent 1X	Cent 1Y	Cent2X	Cent 2Y	Cent 3X	Cent 3Y
0	0,216613	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
0	0,210594	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
0	0,212843	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
0	0,26113	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,275501	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,210594	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,187192	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,264138	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,261772	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060
1	0,279388	-1,4470	-0,0300	-3,4470	-0,0956	-5,4470	-0,0060

Dist,C1	Dist, C2	Dist, C3	Cluster V	MATCH	N_C1X	N_C1Y	N_C2X	N_C2Y	N_C3X	N_C3Y
1,6936	3,7592	5,6696	1,6936	1	0,6	0,237977	0	0	0	0
1,6876	3,7532	5,6636	1,6876	1						
1,6898	3,7554	5,6658	1,6898	1						
1,7381	3,8037	5,7141	1,7381	1						
2,7525	4,8181	6,7285	2,7525	1						
2,6876	4,7532	6,6636	2,6876	1						
2,6642	4,7298	6,6402	2,6642	1						
2,7411	4,8067	6,7171	2,7411	1						
2,7388	4,8044	6,7148	2,7388	1						
2,7564	4,8220	6,7324	2,7564	1						

Pengujian Manual Manhattan Distance Iterasi 1

Tahun	Gini Rasio	Cent 1X	Cent 1Y	Cent2X	Cent 2Y	Cent 3X	Cent 3Y
0	0,216613	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
0	0,210594	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
0	0,212843	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
0	0,26113	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,275501	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,210594	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,187192	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,264138	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,261772	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,279388	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000

Dist,C1	Dist, C2	Dist, C3	Cluster V	MATCH	N_C1X	N_C1Y	N_C2X	N_C2Y	N_C3X	N_C3Y
0,6214	0,2166	0,2166	0,2166	2	1	0,246431	0	0,225295	0	0
0,6274	0,2106	0,2106	0,2106	2						
0,6251	0,2128	0,2128	0,2128	2						
0,6232	0,2611	0,2611	0,2611	2						
0,4375	1,2755	1,2755	0,4375	1						
0,4274	1,2106	1,2106	0,4274	1						
0,4508	1,1872	1,1872	0,4508	1						
0,4262	1,2641	1,2641	0,4262	1						
0,4238	1,2618	1,2618	0,4238	1						
0,4414	1,2794	1,2794	0,4414	1						

Pengujian Manual Manhattan Distance Iterasi 2

Tahun	Gini Rasio	Cent 1X	Cent 1Y	Cent2X	Cent 2Y	Cent 3X	Cent 3Y
0	0,216613	1,0000	0,2464	0,0000	0,2253	0,0000	0,0000
0	0,210594	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
0	0,212843	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
0	0,26113	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,275501	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,210594	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,187192	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,264138	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,261772	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000
1	0,279388	0,6000	0,2380	0,0000	0,0000	0,0000	0,0000

Dist,C1	Dist, C2	Dist, C3	Cluster V	MATCH	N_C1X	N_C1Y	N_C2X	N_C2Y	N_C3X	N_C3Y
1,0298	0,0087	0,2166	0,0087	2	1	0,246431	0	0,225295	0	0
1,0358	0,0147	0,2106	0,0147	2						
1,0336	0,0125	0,2128	0,0125	2						
1,0147	0,0358	0,2611	0,0358	2						
0,0291	1,0502	1,2755	0,0291	1						
0,0358	1,0147	1,2106	0,0358	1						
0,0592	1,0381	1,1872	0,0592	1						
0,0177	1,0388	1,2641	0,0177	1						
0,0153	1,0365	1,2618	0,0153	1						
0,0330	1,0541	1,2794	0,0330	1						

.

.

...