PENGELOMPOKAN DATA PEMINJAMAN BUKU DI PERPUSTAKAAN SMA 5 MUHAMMADIYAH DENGAN MENGGUNAKAN METODE *K-MEANS CLUSTERING*

SKRIPSI

Disusun untuk memenuhi sebagian persyaratan
mencapai derajat Sarjana

Disusun Oleh:

MUHAMMAD SULTAN REZA ADITYA NURRAHMAN
2000018182

PROGRAM STUDI S1 INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS AHMAD DAHLAN

2024

PENGELOMPOKAN DATA PEMINJAMAN BUKU DI PERPUSTAKAAN SMA 5 MUHAMMADIYAH DENGAN MENGGUNAKAN METODE *K-MEANS CLUSTERING*

SKRIPSI

Disusun Oleh:

MUHAMMAD SULTAN REZA ADITYA NURRAHMAN
2000018182

PROGRAM STUDI S1 INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS AHMAD DAHLAN

2024

LEMBAR PERSETUJUAN PEMBIMBING

SKRIPSI

PENGELOMPOKAN DATA PEMINJAMAN BUKU DIPERPUSTAKAAN SMA 5 MUHAMMADIYAH DENGAN MENGGUNAKAN METODE *K-MEANS CLUSTERING*

Dipersiapkan dan disusun oleh:

MUHAMMAD SULTAN REZA ADITYA NURRAHMAN
2000018182

Program Studi S1 Informatika
Fakultas Teknologi Industri
Universitas Ahmad Dahlan

Telah disetujui oleh:

Pembimbing

Lisna Zahrotun, S.T., M.Cs.

19840911 200909 011 1058758

Dipindai dengan CamScanner

LEMBAR PENGESAHAN

SKRIPSI

PENGELOMPOKAN DATA PEMINJAMAN BUKU DI PERPUSTAKAAN SMA 5 MUHAMMADIYAH DENGAN MENGGUNAKAN METODE *K-MEANS CLUSTERING*

Dipersiapkan dan disusun oleh:

MUHAMMAD SULTAN REZA ADITYA NURRAHMAN
2000018182

Telah dipertahankan di depan Dewan Penguji
pada 19 Juni 2024
dan dinyatakan telah memenuhi syarat

Susunan Dewan Penguji

Ketua	**: Lisna Zahrotun, S.T., M.Cs.**	21-6-2024
Penguji 1	**: Dr. Ardiansyah, S.T., M.Cs.**	
Penguji 2	**: Murein Miksa Mardhia, S.T., M.T.**	21/6/24

Yogyakarta, 24 Juni 2024
Dekan Fakultas Teknologi Industri
Universitas Ahmad Dahlan

Prof. Dr. Ir. Siti Jamilatun, M.T.
19660812 199601 011 0784324

Dipindai dengan CamScanner

Pernyataan Persetujuan Akses

Saya yang bertanda tangan di bawah ini:

Nama	:	Muhammad Sultan Reza Aditya Nurrahman
NIM	:	2000018182
Email	:	Muhammad2000018182@webmail.uad.ac.id
Program Studi	:	S1 Informatika
Fakultas	:	Teknologi Industri
Judul Tesis	:	Pengelompokan Data Peminjaman Buku Di Perpustakaan SMA 5 Muhammadiyah Dengan Menggunakan Metode *K-means Clustering*

Dengan ini Saya menyerahkan hak sepenuhnya kepada Perpustakaan Universitas Ahmad Dahlan untuk menyimpan, mengatur akses serta melakukan pengelolaan terhadap karya saya ini dengan mengacu pada ketentuan akses tesis elektronik sebagai berikut (beri tanda pada kotak):

☑ Saya **(mengijinkan/tidak mengijinkan)*** karya tersebut diunggah ke dalam aplikasi Repository Perpustakaan Universitas Ahmad Dahlan.

Demikian pernyataan ini Saya buat dengan sebenarnya.

Yogyakarta, 19 Juni 2024

Yang Menyatakan

Muhammad Sultan Reza A.N.

Mengetahui,

Dosen Pembimbing Skripsi

Lisna Zahrotun, S.T., M.Cs.

Dipindai dengan CamScanner

LEMBAR PERNYATAAN KEASLIAN

SURAT PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : Muhammad Sultan Reza Aditya Nurrahman

NIM : 2000018182

Prodi : Informatika

Judul TA/Skripsi :Pengelompokan Data Peminjaman Buku Di Perpustakaan SMA 5 Muhammadiyah Dengan Menggunakan Metode *K-means Clustering*

Dengan ini saya menyatakan bahwa Laporan Tugas Akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar Ahli Madya/Kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 21 Juni 2024

Mengetahui,
Dosen Pembimbing

Lisna Zahrotun, S.T., M.Cs.
19840911 200909 011 1058758

Yang menyatakan,

Muhammad Sultan Reza A.N.
2000018182

Dipindai dengan CamScanner

KATA PENGANTAR

Puji dan syukur saya panjatkan atas kehadirat Allah S.W.T, atas limpahan rahmat dan hidayah-Nya penulis dapat menyelesaikan proposal penelitian yang berjudul "Pengelompokan Data Peminjaman Buku Diperpustakaan Sma 5 Muhammadiyah Dengan Menggunakan Metode *K-means Clustering*" sebagai syarat kelulusan. Dalam penyusunan laporan ini tentunya tidak terlepas dari arahan serta bimbingan dari berbagai pihak. Oleh karena itu penulis mengucapkan terimakasih kepada :

1. Bapak Prof. Dr. Muchlas, M.T. selaku rektor Universitas Ahmad Dahlan.
2. Prof. Dr. Ir. Siti Jamilatun, M.T. selaku Dekan Fakultas Teknologi Indutsri Universitas Ahmad Dahlan.
3. Bapak Dr. Murinto, S.Si., M.Kom. selaku Kepala Program Studi Informatika Universitas Ahmad Dahlan dan Dosen Pembimbing Akademik yang senantiasa memberikan arahan, dorongan dan motivasi.
4. Ibu Lisna Zahrotun, S.T., M.Cs. selaku dosen Pembimbing yang senantiasa sabar membimbing dan memberikan arahan.
5. Bapak dan Ibu Dosen Program Studi Informatika Universitas Ahmad Dahlan, yang telah memberikan dorongan dalam perencanaan dan penyelesaian laporan ini.

Penulis menyadari bahwa laporan ini masih terdapat banyak kekurangan, oleh karena itu penulis mengharapkan kritik dan saran yang membagun agar laporan ini menjadi lebih baik.

Yogyakarta, Juli 2024

Muhammad Sultan Reza A.N.

Dipindai dengan CamScanner

MOTTO

"Ketahuilah bahwa kemenangan bersama kesabaran, kelapangan bersama kesempitan, dan kesulitan bersama kemudahan." -HR. Tirmidzi

"Keberhasilan adalah perjalanan panjang dari satu kegagalan ke kegagalan berikutnya tanpa kehilangan semangat." - Winston Churchill

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE PROGRAM

DAFTAR SINGKATAN DAN ARTI LAMBANG

DBI = *Davies-Bouldin Score*

CHI = Calinski-Harabasz Index

SMA = Sekolah Menengah Atas

KDD = Knowledge Discovery in Database

CD = Cleaning Data

Max = Maximal

Min = Minimal

Me = Median

RPL = Rekayasa Perangkat Lunak

Dipindai dengan CamScanner

ABSTRAK

Penurunan jumlah peminjaman buku di Perpustakaan SMA 5 Muhammadiyah sejak pandemi COVID-19 telah menjadi masalah signifikan. Meskipun telah dilakukan pengadaan buku baru, upaya ini belum cukup efektif dalam meningkatkan minat peminjaman. Oleh karena itu, diperlukan pengolahan data peminjaman buku sebagai Upaya untuk menaikan minat baca. Penelitian ini menggunakan metode K-Means Clustering untuk membantu perpustakaan menganalisis dan merencanakan pengadaan buku yang lebih efisien dan tepat sasaran.

Proses pengelompokan data akan menggunakan metode K-Means Clustering. Dengan dimulai dari tahap pertama adalah pengolahan data, yang mencakup pengumpulan data peminjaman buku dan pra-pemrosesan data, termasuk tahap importing library, reading data, cleaning, selecting, dan transforming data. Transformasi data dilakukan dengan menggabungkan antar judul, genre, dan penulis menjadi satu keterangan dan membuat ID baru berdasarkan keunikan dari ketiga data tersebut. Setelah itu, dilakukan tahap normalisasi data untuk memastikan semua fitur berada dalam skala yang sama. Jumlah klaster yang optimal ditentukan menggunakan metrik evaluasi seperti Silhouette Score, Davies-Bouldin Score, dan Calinski-Harabasz Score. Setelah jumlah klaster ditentukan, algoritma K-Means mengelompokan data ke dalam beberapa klaster berdasarkan karakteristik peminjaman buku. Data yang telah dikelompokkan kemudian divisualisasikan menggunakan bar chart untuk mempermudah proses representasi pengetahuan dengan melihat karakteristik tiap klaster.

Hasil pengelompokan data peminjaman buku menggunakan 3 klaster terbukti optimal, dengan Silhouette Score mencapai 0,791. Skor ini mendekati 1, menandakan kualitas klaster yang sangat baik. Selain itu, tingkat kestabilan 3 klaster ini juga sangat baik, dibuktikan dengan metriks Davies-Bouldin Score yang menghasilkan score sebesar 0,626 dan Calinski-Harabasz Score sebesar 615,093. Adapun untuk pengelompokan menggunakan 3 cluster menghasilkan cluster dengan buku paling diminati dengan buku didominasi genre religi, cukup diminati dengan buku didominasi genre drama dan romance, dan kurang diminati didominasi buku bergenre drama dan Umum. Dari hasil ini diharapkan dapat memudahkan perpustakaan dalam merencanakan pengadaan buku yang lebih efektif dan efisien.

Keyword : Data Mining; Data Perpustakaan; KMeans Clustering;

BAB I. Pendahuluan

1.1 Latar Belakang

Secara umum Perpustakaan adalah tempat yang menyediakani sarana sumber informasi dan ilmu pengetahuan untuk menyimpan bahan pustaka yang dipakai oleh pemakai untuk menggali ilmu sumber informasi, termasuk buku, jurnal, dan bahan bacaan lain[1]. Perpustakaan berperan sebagai pusat informasi dan edukasi yang mendukung bagi kegiatan literasi, penelitian, dan pembelajaran bagi berbagai kalangan, mulai dari pelajar, mahasiswa, hingga masyarakat umum. Oleh karena itu, perpustakaan memiliki peran yang sangat vital dalam pengembangan ilmu pengetahuan dan budaya membaca. Begitu pula dengan perpustakaan SMA 5 Muhammadiyah yang merupakan salah satu fasilitas pendidikan penting di sekolah tersebut. Para siswa menjadikan Perpustakaan ini sebagai sumber bacaan dan sarana bagi mereka untuk memperoleh pengetahuan mengenai buku yang diminati. Namun hal tersebut mulai memudar semenjak adanya persebaran virus Covid 19 yang menyebabkan siswa belajar dari rumah. Pembelajaran ini dilakukan secara daring menggunakan google meet dan sebagainya[2]. Pembelajaran yang dilakukan melalui perangkat seluler menjadikan siswa kurang tertarik membaca buku. Kebiasaan ini terus berlanjut bahkan setelah pembelajaran dilanjutkan secara offline. Dengan demikian terjadi penurunan yang sangat drastis pada data peminjaman buku perpustakaan dari tahun 2018 – 2019 dan 2021 - 2023

Pihak Perpustakaan SMA 5 Muhammadiyah menyadari mengenai hal ini, berbagai upaya seperti peningkatan jumlah buku sudah mereka lakukan untuk meningkatkan minat baca para siswa, akan tetapi usaha tersebut masih kurang membuahkan hasil. Setelah dilakukannya wawancara bersama perpustakaan, diketahui bahwa pihak perpustakaan memiliki kurangnya pengetahuan dalam melakukan pengolahan dan analisis untuk data peminjaman buku.

Pengolahan dan analisis data ini tidak hanya penting untuk administrasi dan pengelolaan koleksi, tetapi juga untuk memahami kebutuhan dan preferensi pengguna. Analisis data peminjaman buku sangat diperlukan untuk memahami pola peminjaman buku diperpustakaan. Oleh karenanya pengadaan buku ini dianggap masih kurang efektif, karena belum diketahui pola peminjaman buku dari rentang tahun 2018 - 2023 .

Proses pengolahan dan analisis data perpustakaan dapat menggunakan berbagai cara, salah satunya adalah data mining. Secara sederhana Data Mining adalah proses untuk menambang atau menggali informasi yang tersembunyi dari bongkahan data besar. Inti dari proses-proses Knowledge Discovery in Database (KDD) adalah Data mining, dengan algoritma yang mengeksplor dan membangun model data[3]. Dengan adanya ilmu data mining beserta dengan berbagai metodenya bisa dimanfaatkan untuk ekstrasi data berskala kecil sampai besar agar bisa menghasilkan informasi dan pengetahuan yang bermanfaat. Dalam memutuskan suatu metode data mining yang akan digunakan perlu memahami ciri khas dari setiap metode yang berhubungan dengan pengelompokan data perpusatakaan.

Ada berbagai metode yang sering digunakan dalam penelitian pengelompokan data namun memiliki kondisi penggunaan yang berbeda beda berdasarkan kelebihan dan kekurangannya seperti metode Density-Based Spatial Clustering of Applications with Noise(DBSCAN) yang efektif dalam mengidentifikasi kelompok dengan bentuk dan ukuran yang tidak teratur namun memiliki kekurangan dalam kerumitan mencari parameter radius dari suatu titik dan minimum jumlah titik untuk dianggap sebagai cluster dan kurang cocok digunakan untuk data dengan kepadatan tidak seragam pada pola peminjaman buku sehingga dapat menghasilkan ketidak konsistenan hasil. Kemudian *K-Means Clustering* merupakan metode yang paling sederhana, mudah dipahami, dan efisien dalam melakukan pengelompokan data namun sensitif terhadap Pemilihan Jumlah Kelompok (K) karena itu perlu

dibantu dengan matriks evaluasi seperti *silhouette coefficient*. Serta, metode *Fuzzy C-Means* yang meskipun memiliki keunggulan representasi pola yang lebih kompleks dibandingkan *K-Means Clustering* namun diperlukan pengetahuan yang cukup mendalam sehingga menyulitkan pihak perpustakaan memahami tentang konsep matematika, komputasi pemrograman, dan analisis data parameter fuzziness(m) yang menentukan besaran partisi data karena setiap titik data dapat berada di lebih dari 1 kelompok. [4].

Selain dari 3 metode tersebut ada berbagai metode lainnya seperti Model-based clustering, dan Hierarchical clustering yang memiliki keunggulan dibandingkan metode lainnya, namun jarang sekali digunakan dalam penelitian pengelompokan data perpustakaan karena memiliki kekurangan pada proses komputasi yang bergantung pada distribusi data yang tepat dan rawan terjadi kesalahan pada hasil yang diperoleh. Pada kasus dimana pihak perpustakaan memiliki kurangnya pengetahuan mengenai pengolahan data buku, maka berdasarkan karakteristiknya metode paling cocok adalah K-Means Clustering. J.Han, dkk dalam bukunya [4] menyatakan bahwa *K-Means Clustering* adalah metode pengelompokan (*clustering*) yang paling sederhana, mudah dipahami, mudah diimplementasikan, dan komputasi yang efisien terhadap data yang berukuran cukup besar, sehingga membantu mempermudah pihak perpustakaan dalam menganalisis pola peminjaman buku. Untuk semakin meyakinkan digunakannya metode *K-Means Clustering* pada penelitian ini, perlu juga melakukan observasi pada keberhasilan penelitian terdahulu dengan menggunakan kasus dan metode serupa.

Penelitian tahun 2021 oleh Januardi Nasir mengenai penerapan data mining clustering dalam mengelompokan buku dengan metode k-means untuk tujuan membantu pihak perpustakaan dalam mengetahui kelompok/cluster buku mana yang sering dipinjam menggunakan kmeans clustering untuk menemukan pola peminjaman buku terbukti efektif memberikan kemudahan bagi pihak perpustakaan dalam melakukan kegiatan pengadaan dan

pengelolaan buku diperpustakaan[5].

Penelitian yang dilakukan oleh Baker tahun 2020 mengenai implementasi data mining dalam menentukan penambahan koleksi buku di perpustakaan menggunakan algoritma Kmeans clustering dengan tujuan membantu pihak perpustakaan memantau jumlah buku yang sering dipinjam setiap bulan. Perhitungan Kmeans Clustering ini menghasilkan cluster optimal di mana terdapat dua jenis klasifikasi buku yang banyak dipinjam yang termasuk dalam cluster-1, sementara cluster-2 berisi 51 jenis klasifikasi buku yang kurang atau jarang dipinjam. Melalui penerapan data mining menggunakan algoritma K-Means Clustering, diperoleh informasi yang dapat membantu pihak perpustakaan dalam mengidentifikasi buku-buku yang paling sering dipinjam. Hal ini memungkinkan pihak perpustakaan untuk memantau jumlah buku yang sering dipinjam setiap bulan,dan membantu dalam menentukan penambahan koleksi buku secara lebih efektif dan efisien[6].

Berdasarkan keberhasilan penelitian terdahulu dan kecocokan antara karakteristik *K-Means Clustering* dengan akar masalah penelitian, maka diputuskan penggunaan metode *K-Means Clustering* dalam penelitian ini. Data peminjaman buku akan diolah menjadi 3 cluster diminati, cukup diminati, dan kurang diminati berdasarkan karakteristik yang sama untuk memudahkan pihak Perpustakaan dalam analisis. Berdasarkan pemaparan latar belakang diatas maka diusulkan penelitian ini untuk mengidentifikasi pola peminjaman buku di perpustakaan SMA 5 Muhammadiyah

1.2 Batasan Masalah

Peninjauan Batasan masalah ini penting agar penelitian ini tidak terlalu luas dan agar memiliki pembatas untuk ditinjau:

1. Menggunakan data peminjaman buku SMA 5 Muhammadiyah dengan rentang tahun 2018-2023.
2. Mengidentifikasi pola peminjaman buku siswa di perpustakaan SMA 5 Muhammadiyah berdasar atribut judul buku, genre buku, penulis buku, dan frekuensi peminjaman-nya tiap tahun
3. Penelitian akan melakukan pengolahan data peminjaman buku menggunakan *K-Means CLustering* untuk membantu pihak perpustakaan dalam pengadaan dan pengelolaan koleksi buku berdasarkan hasil pembagian data menjadi beberapa cluster berdasarkan karakteristiknya.
4. Evaluasi cluster menggunakan matriks silhouette coefficient , davied bouldin score, dan Calinski-Harabasz Index

1.3 Rumusan Masalah

1. Penurunan signifikan dalam jumlah peminjaman buku terjadi di SMA 5 Muhammadiyah pasca pandemi COVID-19, menimbulkan kebutuhan akan pengadaan buku dengan langkah yang strategis.
2. Kurangnya pengetahuan dan keterampilan dalam pengolahan data di perpustakaan SMA 5 Muhammadiyah menghambat identifikasi dan respons terhadap pola peminjaman buku.

1.4 Tujuan Penelitian

Tujuan penelitian adalah :

1. Mencari pola peminjaman buku menggunakan metode K-Means Clustering
2. Melakukan evaluasi kualitas cluster berdasarkan matriks evaluasi
3. Melakukan analisis langkah strategis untuk direkomendasikan kepada pihak perpustakaan berdasarkan hasil analisis pola yang ditemukan

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah :

1. Membantu pengurus perpustakaan dalam menemukan pola peminjaman buku menggunakan metode K-Means Clustering
2. Membantu pengurus dalam pengadaan buku perpustakaan berdasarkan analisis pola peminjaman dari hasil kelompok cluster yang sudah dibuat.

BAB II. Tinjauan Pustaka

2.1 Kajian Terdahulu

Tinjauan pustaka merupakan analisis terhadap berbagai penelitian sebelumnya yang berfungsi sebagai panduan untuk penelitian yang akan datang. Selain membahas penelitian terdahulu, bab ini juga akan memaparkan teori-teori yang mendukung penelitian ini. Berikut adalah beberapa ringkasan dari hasil penelitian terdahulu yang dapat digunakan sebagai referensi untuk penelitian yang akan dilakukan oleh penulis.

1. Haryani, dkk tahun 2021 [3] melakukan penelitian membantu perpustakaan Yayasan Nurul Islam Indonesia yang hanya memiliki ±2000 judul buku dalam mengatur koleksi buku.Penelitian ini mengelompokan data kategori buku menjadi tiga cluster yaitu paling diminati, diminati dan kurang diminati. Atribut yang digunakan dalam pengolahan data meliputi jumlah peminjam buku dan jumlah stok buku. Hasil yang didapatkan untuk data kategori buku yang berada pada cluster paling diminati, inilah nantinya yang akan dijadikan sebagai bahan evaluasi bagi pihak pustakawan dalam meningkatkan koleksi buku diperpustakaan Yayasan Nurul Islam Indonesia Baru.

2. Penelitian tahun 2020 [7] dengan tujuan untuk mengetahui hasil dari pengelompokan penjualan berdasarkan jumlah penjualan per bulan selama 5 tahun menggunakan metode Kmeans Clustering. Objek penelitian yang penulis teliti berasal dari 900 data yang diperoleh dari gaikindo. . Data Penjualan nantinya akan dikelompokan berdasarkan kemiripan data tersebut sehingga data dengan karakteristik yang sama akan berada dalam satu cluster. Atribut yang digunakan adalah brand dan penjualan. Cluster yang terbentuk setelah dilakukan proses *K-Means Clustering* terbagi menjadi tiga cluster yaitu Cluster 0 jumlah anggota 235 dengan presentase 26% dikategorikan Laris, *Cluster* 1 jumlah anggota

604 dengan presentase 67% dikategorikan Kurang Laris, dan *Cluster* 2 jumlah angota 61 dengan presentase 7% dikategorikan Paling Laris, dari proses clustering diatas dapat diperoleh validasi DBI (*Davies Bouldin Index*) dengan nilai 0,341.

3. Penelitian tahun 2022 [8] dengan tujuan untuk Untuk Analisa Penjualan Pada Toko Yana Sport. Penelitian dengan metode *K-means Clustering* menetapkan 2 kelompok yaitu laris terjual dan tidak laris terjual. Hasil Data Performance menejelaskan bahwa Cluster 0 dengan nilai 110 dikategorikan sebagai Tidak Laris Terjual sedangkan 21389 dikategorikan sebagai laris terjual. Hasil rekomendasai penelitian ini mendapatkan informasi atau pola dari penerapan algoritma k-means dengan data penjualan terdapat sebanyak 99 item barang yang laris terjual dan terdapat 23 item barang yang tidak terjual sehingga pemilik dapat melakukan strategi penjualan dan pembelian ulang berdasarkan barang yang laris terjual.

4. Penelitian tahun 2021 [9] dengan tujuan membuat sebuah aplikasi untuk menghasilkan informasi berupa pengelompokan data pengunjung untuk mengetahui jenis pengunjung apa saja yang sering berkunjung, dan data peminjam digunanakan untuk mengetahui buku apa saja yang banyak dipinjam menggunakan metode *K-Means*, yang mana digolongkan menjadi 3 cluster yaitu *cluster* 1 (sangat sering dipinjam), *cluster* 2 (sering dipinjam), *cluster* 3 (jarang dipinjam). Variabel yang digunakan untuk mendapatkan hasil perhitungan K-Means peminjam yaitu total kemunculan nama buku pada setiap bulannya dalam priode setahun dan variabel yang digunakan pada K-Means pengunjung yaitu total kemunculan nama jurusan atau jenis pengunjung pada setiap bulannya dalam priode setahun. Hasil dari pengelompokan tersebut akan digunakan pihak Perpustakaan untuk membantu proses evaluasi layanan kunjungan perpustakaan dalam memberikan saran untuk pengadaan koleksi buku serta mengevaluasi koleksi baik buku maupun dokumentasi yang

banyak dipinjam, dan juga dapat digunakan untuk mapping tata letak buku yang sering dipinjam.

5. Penelitian tahun 2022 [10] dengan tujuan menentukan strategi promosi berdasarkan daerah asal dari profil mahasiswa baru yang mendaftar. Metode yang digunakan adalah *Clustering* dan algoritma yang digunakan adalah *K-Means*. Data yang sudah diolah akan dilakukan proses analisa menggunakan teknik *Knowledge Discovery in Databases* (KDD) dengan lima tahapan yaitu *selection, preprocessing, transformation, data mining dan evaluation*. Proses implementasi dalam penelitian ini menggunakan software WEKA versi 3.8.5, dan menghasilkan tiga klasterisasi yaitu Cluster 1 berjumlah 189 data dengan presentase 42%, Cluster 2 berjumlah 186 data dengan presentase 41% dan Cluster 3 berjumlah 78 data dengan presentase 17%. Dengan kecepatan komputasinya sebesar 0.01 detik. Dari hasil Cluster ini akan ditentukan langkah promosi yang akan digunakan.

Tabel 2. 1 *Kajian Terdahulu*

Peneliti	Dataset	Variabel	Metode	Hasil
Haryani , Dicky Nofriansyah,dan Ita Mariami [3]	Data Perpustakaan Yayasan Nurul Islam Indonesia	Kategori buku, peminjam buku, dan stock buku	Kmeans Clustering	Pengelompokan buku berdasarkan data yang diperoleh dari Perpustakaan Yayasan Nurul Islam Indonesia Baru untuk pusat cluster pertama dinyatakan ada 10 kategori buku yang paling diminati yang dimana dapat dilihat dari jumlah peminjam buku lebih besar peminatnya sedangkan jumlah stok kategori buku lebih sedikit. Untuk pusat cluster kedua dinyatakan ada 6 kategori buku yang diminati yang dimana dilihat dari jumlah peminjam buku peminatnya sama dengan jumlah stok buku yang tersedia. dan pada pusat cluster ketiga dinyatakan ada 14 kategori buku yang kurang diminati ini dilihat dari jumlah peminjam buku peminatnya lebih sedikit jumlahnya sedangkan jumlah stoknya lebih besar.
Sufajar Butsianto , dan Nindi Tya Mayangwulan [7]	Data Penjualan Mobil Gaikindo tahun 2015 – 2019	bulan, brand, dan penjualan	Kmeans Clustering	Cluster yang terbentuk setelah dilakukan proses K-Means Clustering terbagi menjadi tiga cluster yaitu Cluster 0 jumlah anggota 235 dengan presentase 26% dikategorikan Laris, Cluster 1 jumlah anggota 604 dengan presentase 67% dikategorikan Kurang Laris, dan Cluster 2 jumlah angota 61 dengan presentase 7% dikategorikan Paling Laris, dari proses clustering diatas dapat diperoleh validasi DBI (Davies Bouldin Index) dengan nilai 0,341
Agung Nugraha, Odi Nurdiawan , dan Gifthera Dwilestari [8]	Data transaksi penjualan toko Yana yang sport	Nama barang, Kode barang, Stock awal, Stock keluar, stock akhir	Kmeans Clustering	Hasil rekomendasai penelitian ini kmendapatkan informasi atau pola dari penerapan algoritma k-means dengan data penjualan terdapat sebanyak 99 item barang yang laris terjual dan terdapat 23 item barang yang tidak terjual sehingga pemilik dapat melakukan strategi penjualan dan pembelian ulang berdasarkan barang yang laris terjual.

Andy Febrianto, Sentot Achmadi, Agung Panji Sasmito [9]	Data pengunjung perpustakaan ITN Malang	Judul buku, dan frekuensi peminjaman tiap bulan	Kmeans Clustering	Hasil penelitian ini adalah produk berupa aplikasi website, produk mempunya fitur yaitu sistem pada website dapat memberikan analisis informasi pengunjung dan peminjam, berdasarkan pengujian fungsional sistem seluruhnya berhasil dan berjalan dengan baik, berdasarkan pengujian pengguna diketahui hasil presentase responden 50% Sangat Baik, 48% Baik dan 2% Kurang Baik, berdasarkan penelitian disimpulkan bahwa semua fitur dapat berjalan dengan baik pada browser Mozilla Firefox 83.0 dan Google Chrome 87.0.4280.88, berdasarkan pengujian penguna mayoritas menilai sangat baik terhadap penggunaan aplikasi keuntungan yang didapatkan jika menggunaakan aplikasi ini yaitu dapat memudahkan dalam memberikan seran untuk pengadaan buku secara komputerisasi dan kerugian jika tidak menggunakan aplikasi ini yaitu dalam penentuan untuk pengadaan buka masih secara manual.
Nanda Ayu Rahmalinda dan Arief Jananto [10]	Data penerimaan mahasiswa baru Sekolah Tinggi Ilmu Ekonomi Assholeh Pemalang	Jenis kelamin, asal sekolah, alamat mahasiswa, program study	Kmeans Clustering	Proses implementasi dalam penelitian ini menggunakan software WEKA versi 3.8.5, dan menghasilkan tiga klasterisasi yaitu Cluster 1 berjumlah 189 data dengan presentase 42%, Cluster 2 berjumlah 186 data dengan presentase 41% dan Cluster 3 berjumlah 78 data dengan presentase 17%. Dengan kecepatan komputasinya sebesar 0.01 detik. Proses promosi dilakukan dengan mengirimkan team marketing di persebaran wilayah pada kabupaten yang mendominasi dengan cara mendatangi langsung dan melakukan sosialisasi untuk mengenalkan Sekolah Tinggi Ilmu Ekonomi Assholeh Pemalang dikalangan siswa/i, membagikan brosur dan menempelkan pamflet serta melakukan penyelarasan menggunakan strategi promotion mix.

Dengan melihat pada tabel 2.1 Penelitian ini menawarkan kebaruan dalam pengelompokan data peminjaman buku di perpustakaan SMA 5 Muhammadiyah melalui penggunaan metode K-means clustering yang ditingkatkan dengan integrasi berbagai metrik evaluasi. Inovasi pertama adalah penggunaan kombinasi metrik *Silhouette Score, Davies-Bouldin Score, dan Calinski-Harabasz Score* secara simultan untuk menentukan jumlah klaster yang optimal, memberikan pendekatan yang lebih komprehensif dan akurat. Kemudian untuk data ditransformasikan dengan menggabungkan judul buku, genre, dan penulis menjadi ID agar memudahkan dalam melakukan proses klasterisasi tanpa kehilangan informasi unik judul buku, genre, dan penulis

2.2 Landasan Teori

2.2.1 Pengertian *Data Mining*

Suatu konsep yang dipakai dalam menghasilkan suatu aturan dalam penemuan pengetahuan disebut data mining. *Data mining* atau penambangan data merupakan metode, teknik, artificial intelegent dan mesin pembelajaran yang diekstrasi sehinga menghasilkan suatu pengetahuan dan informasi yang berguna yang tersimpan dalam suatu database. Pada prinsipnya *data mining* mewarisi banyak aspek dan teknik bidang-bidang ilmu. Data mining bukanlah sesuatu hal yang baru, karena data mining merupakan akar dari berbagai bidang ilmu tersebut [5]. *Data mining* sendiri memiliki banyak fungsi. Fungsi utamanya sendiri ada dua; Yakni, fungsi deskriptif dan fungsi prediktif [11].

1. Deskriptif

Fungsi deskriptif dalam data mining adalah fungsi yang bertujuan untuk lebih memahami tentang suatu observasi. Proses tersebut diharapkan dapat mempelajari perilaku dari data. Informasi ini kemudian dapat digunakan untuk menemukan karakteristik data tersebut, dan kemudian dapat menggunakan data mining deskriptif untuk menemukan pola tersembunyi

dalam data. Dengan kata lain, jika pola tersebut berulang dan memiliki nilai, maka karakteristik dari data tersebut diketahui [12].

2. Prediktif

Fungsi prediktif adalah fungsi bagaimana proses nantinya menemukan pola tertentu dalam data. Pola-pola ini dapat dipelajari dari berbagai variabel dalam data [12]. Setelah pola ditemukan, rumus yang dihasilkan dapat digunakan untuk memprediksi variabel lain yang nilai atau jenisnya tidak diketahui, dan karena itu dianggap sebagai fungsi prediksi untuk satu fungsi, tetapi juga melakukan analisis prediksi. Fungsi ini juga dapat digunakan untuk memprediksi variabel tertentu yang tidak diketahui, sehingga fungsi ini memudahkan dan bermanfaat untuk memperbaiki hal-hal penting tersebut bagi mereka yang membutuhkan prediksi yang akurat. Operasi penambangan data lainnya meliputi: karakterisasi, diskriminasi, asosiasi, klasifikasi, pengelompokan, analisis outlier dan tren, dan lain-lain.

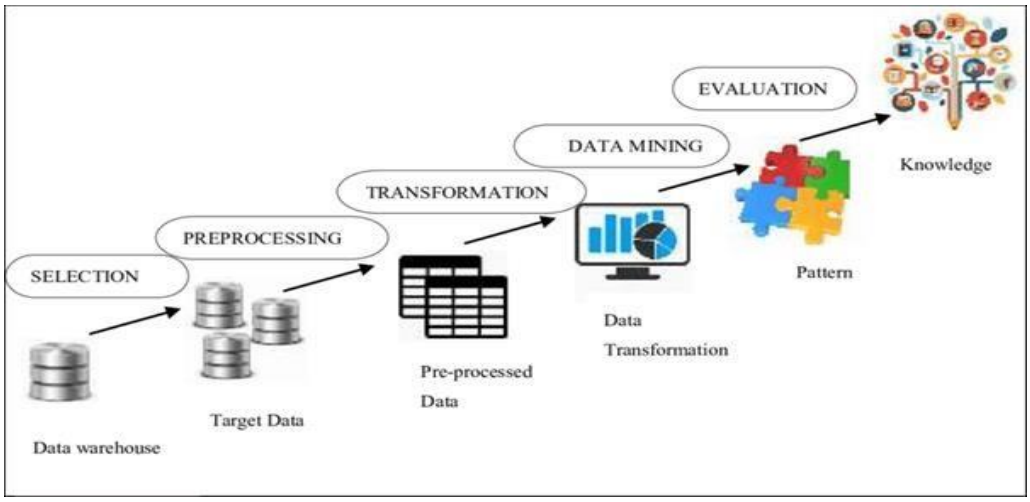
2.2.2 Tahapan *Data Mining*

Tahapan *data mining* adalah pembersihan data, integrasi data, pemilihan data, transformasi, penambangan data, penilaian pola dan representasi data. Suatu proses atau tahapan dimulai dengan data mentah dan diakhiri dengan data atau informasi yang telah diproses [13]. Adapun penjelasan prosesnya adalah sebagai berikut :

1. Pembersihan data, proses mengeluarkan data yang tidak lengkap, tidak benar dan kontradiktif dari pengumpulan data. Untuk mempelajari lebih lanjut tentang pemrosesan data, lihat juga Manajemen Siklus Hidup Data.
2. Integrasi data, proses integrasi data dimana data berulang digabungkan.
3. Seleksi, proses memilih atau memilih informasi yang relevan untuk dianalisis dari data yang tersedia.

4. Transformasi data, proses mentransformasikan data terpilih untuk ditambang dengan menggunakan metode dan data mining.
5. Penambangan data, proses utama menggunakan berbagai teknik untuk mengekstraksi berbagai kemungkinan pola untuk mendapatkan informasi yang berguna.
6. Evolusi Pola, proses mengidentifikasi pola menarik yang ditemukan sebelumnya berdasarkan algoritma yang ditentukan.
7. Presentasi pengetahuan, merupakan tahap akhir dari proses dimana teknik visualisasi digunakan untuk membantu pengguna memahami dan menginterpretasikan hasil penambangan pengetahuan.

Untuk penggambaran tahapan dapat dilihat pada gambar 2.1.



Gambar 2. 1 *Proses Data Mining*

2.2.3 Metode Dalam Data mining

Berbagai teknik digunakan dalam proses penambangan data. [12] Teknik teknis-nya adalah sebagai berikut:

1. Pemodelan prediktif, ada dua teknik yaitu klasifikasi dan prediksi nilai.
2. *Database segmentation*, membagi database menjadi beberapa segmen, cluster atau

record yang sama, dengan analisis link metode yang digunakan untuk membuat link antara record individu atau kelompok record dalam database.

3. Deteksi anomali, teknik untuk mengidentifikasi anomali yang menunjukkan penyimpangan dari ekspektasi yang diketahui sebelumnya.
4. *Nearest Neighbor*, yang merupakan teknik prediksi clustering, merupakan teknik tertua yang digunakan dalam data mining.
5. *Clustering* adalah metode pengelompokan data berdasarkan kriteria masing-masing data.
6. *Decision Tree*, merupakan teknik next generation dimana metode ini merupakan model prediksi yang dapat digambarkan sebagai pohon. Setiap node dalam struktur pohon mewakili pertanyaan yang digunakan untuk mengklasifikasikan data.

2.2.4 K-means Clustering

K-Means Clustering adalah teknik analisis data yang mengelompokan objek data menjadi beberapa kelompok atau cluster berdasarkan kesamaan karakteristiknya. J.Han, dkk dengan buku yang berjudul "*Data Mining Concepts and Techniques*" edisi ketiga tahun 2011 menyatakan bahwa *K-Means Clustering* adalah metode pengelompokan (*clustering*) yang paling sederhana, mudah dipahami ,mudah diimplementasikan , dan komputasi yang efisien terhadap data yang berukuran cukup besar, sehingga membantu mempermudah pihak perpustakaan dalam menganalisis pola peminjaman buku. Dua jenis grup data yang biasa digunakan untuk mengelompokan data, yaitu grup data hierarkis dan grup data non-hierarkis. *K-Means* adalah metode pengelompokan data non-hierarkis yang mencoba membagi data yang ada menjadi satu atau lebih *cluster* [9].

Metode *K-Means Clustering* ini membagi data menjadi *cluster* atau kelompok sehingga data dengan karakteristik yang sama dikelompokkan ke dalam cluster yang sama dan data dengan karakteristik berbeda ke dalam kelompok lain. Tujuan dari informasi pengelompokan ini adalah untuk meminimalkan set fungsi tujuan dalam proses pengelompokan, yang biasanya bertujuan untuk meminimalkan variabilitas dalam-cluster dan memaksimalkan variabilitas antar-*cluster*. Keunggulan clustering adalah pengenalan objek (*recognition*), misalnya dalam bidang image processing, *computer vision* atau robot vision. Selain itu, ini adalah sistem pendukung keputusan dan penambangan data, seperti segmentasi pasar, pemetaan wilayah, manajemen pemasaran, dll [14]. *K-Means data clustering* biasanya dilakukan dengan algoritma dasar sebagai Berikut [15] :

1. Tentukan jumlah cluster yang diinginkan
2. Inisialisasi k pusat klaster (centroid) secara random/ acak
3. Tempatkan setiap data atau objek ke klaster terdekat. Kedekatan dua objek ditentukan berdasar jarak. Jarak yang dipakai pada algoritma k-Means adalah Euclidean distance (d). Rata-rata semua data ke akurasi centroid/rata-rata terdekat. Untuk rumus Euclidean distance adalah sebagai persamaan 1 [15] :

$$De\ (x, y) = \sum_{i=0}^{n} (\mathrm{xi} - \mathrm{yi})^2 \qquad (1)$$

Dimana :

1. d(x,y) adalah Euclidean distance antara titik x dan y
2. xi dan yi adalah koordinat ke-I dari titik x dan y masing masing
3. N adalah jumlah dimensi

4. Hitung kembali pusat klaster dengan keanggotaan klaster yang sekarang. Pusat klaster adalah rata-rata (*mean*) dari semua data atau objek dalam klaster tertentu.

Karakteristik dari *K-Means Cluster* [16]**:**

1. Cepat dalam proses clustering
2. Sensitif terhadap nilai centroid
3. Hasil dari *K-Means* adalah cluster yang bersifat *exclusif*

Kekurangan dari *K-Means clustering* [16]:

1. Sensitif terhadap pusat cluster
2. Memerlukan matriks evaluasi seperti silhouette untuk menentukan jumlah cluster yang tepat
3. Kurang mampu menangani data dengan varian berbeda (misalkan membandingkan frekuensi peminjaman dengan jumlah eksemplar buku)

2.2.5 Sillhouette Coefficient

Silhouette Coefficient adalah metrik yang digunakan untuk mengukur seberapa baik objek-objek dalam suatu klasterisasi dikelompokkan. Nilai silhouette untuk setiap objek menunjukkan seberapa mirip objek tersebut dengan klaster tempatnya berada dibandingkan dengan klaster lainnya. Metrik ini dikembangkan oleh Peter J. Rousseeuw dan diperkenalkan dalam makalahnya yang berjudul "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis" pada tahun 1986. Silhouette menjadi alat penting dalam analisis klaster untuk menilai kualitas klasterisasi.

Tujuan dari metrik silhouette adalah memberikan cara objektif untuk mengevaluasi seberapa baik klasterisasi yang dilakukan. Dengan menggunakan nilai silhouette, diketahui

apakah objek-objek dalam klaster benar-benar mirip satu sama lain dan berbeda dari objek-objek di klaster lain. Proses perhitungan silhouette secara manual melibatkan penghitungan rata-rata jarak antara satu objek dengan semua objek dalam klasternya (a(i)) dan rata-rata jarak antara satu objek dengan semua objek di klaster terdekat lainnya (b(i)), kemudian nilai silhouette dihitung dengan menggunakan persamaan 2:

$$S(i) = \sum_{i=0}^{n} \frac{(\mathrm{b(i)} - \mathrm{a(i)})}{(\max(\mathrm{a(i)}, \mathrm{b(i)})} \qquad (2)$$

Keterangan :

1. S adalah Sillhouette score
2. a(i) adalah jarak data dengan cluster a (untuk a dan b bisa menggunakan cluster bebas)
3. b(i) adalah jarak data dengan cluster b

Dalam konteks k-means clustering, metrik silhouette sangat bermanfaat karena membantu dalam memilih jumlah klaster yang optimal dan menilai apakah hasil klasterisasi sudah sesuai. Nilai silhouette yang mendekati 1 menunjukkan klasterisasi yang baik, nilai sekitar 0 menunjukkan tumpang tindih antar klaster, dan nilai negatif menunjukkan bahwa objek mungkin lebih cocok di klaster lain. Rentang nilai -1 hingga 1 dipilih karena memberikan batasan yang jelas dan memungkinkan interpretasi yang intuitif: nilai positif yang tinggi mengindikasikan klaster yang baik, nilai sekitar 0 menunjukkan bahwa objek berada di batas antar klaster, dan nilai negatif menunjukkan bahwa objek lebih cocok untuk klaster yang berbeda. Rentang ini memastikan bahwa hasil klasterisasi bisa dibandingkan secara konsisten di berbagai aplikasi dan jenis data.Pemilihan jumlah klaster optimal melibatkan maksimalkan nilai silhouette rata-rata, yang pada gilirannya membantu mengidentifikasi struktur klaster yang lebih baik dan lebih representatif dalam data yang diberikan [17].

2.2.6 Davies-Bouldin Score dan Calinski-Harabasz Index

Davies-Bouldin Score adalah metrik yang digunakan untuk mengevaluasi kualitas klaster yang dihasilkan oleh algoritma klasterisasi seperti *K-means*. Skor ini dihitung berdasarkan rasio antara jarak dalam klaster *(intra-cluster distance)* dan jarak antar klaster *(inter-cluster distance).* Semakin kecil nilai *Davies-Bouldin Score*, semakin baik klaster yang dihasilkan, karena hal ini menunjukkan bahwa klaster yang terbentuk memiliki kepadatan yang tinggi dan terpisah dengan baik dari klaster lainnya. Metrik ini sangat berguna untuk menentukan seberapa baik klaster yang dihasilkan tanpa memerlukan label data yang sebenarnya, memungkinkan untuk secara obyektif mengevaluasi performa klasterisasi. Untuk perhitungan DBI manual dapat dilihat pada persamaan 3 :

$$DBI = \sum_{i=0}^{n} \frac{((\mathrm{m(a)} - \mathrm{Rm(b)2})}{Nm} \qquad (3)$$

Keterangan :

1. DBI adalah *Davies-Bouldin Score*
2. m(a) adalah Pusat cluster A
3. m(b) adalah Pusat cluster B
4. Nm adalah Jumlah pusat cluster

Calinski-Harabasz Index, juga dikenal sebagai *Variance Ratio Criterion*, adalah metrik lain yang digunakan untuk mengevaluasi kualitas klaster. Indeks ini dihitung berdasarkan rasio antara jumlah dispersi antar klaster dan jumlah dispersi dalam klaster. Semakin tinggi nilai *Calinski-Harabasz Index*, semakin baik klaster yang dihasilkan, karena ini menunjukkan bahwa klaster-klaster tersebut tersebar dengan baik (*inter-cluster dispersion*) dan memiliki variabilitas rendah dalam klaster (*intra-cluster dispersion*). Dengan kata lain, indeks ini mengukur seberapa

kompak dan terpisah klaster-klaster tersebut, memberikan indikasi yang jelas tentang efisiensi klasterisasi. Untuk perhitungan CHI manual dapat dilihat pada persamaan 4 :

$$DBI = \frac{\frac{5\ x \sum_{i=0}^{n} \mathrm{m(i)}^2}{Nm}}{\frac{\sum_{i=0}^{n} m(i)^2}{Nm}} \qquad (4)$$

Keterangan :

1. m(i) adalah Pusat cluster dijumlahkan
2. Nm adalah total jumlah Cluster

Dalam konteks *K-means clustering, Davies-Bouldin Score dan Calinski-Harabasz Index* digunakan untuk menilai dan menentukan kualitas klaster yang dihasilkan. Dengan menggunakan *Davies-Bouldin Score* dan *Calinski-Harabasz Index* akan mengevaluasi apakah jumlah klaster yang dipilih sudah optimal. Misalnya, nilai *Davies-Bouldin Score* yang rendah atau *Calinski-Harabasz Index* yang tinggi untuk suatu jumlah klaster tertentu menunjukkan bahwa klaster tersebut memiliki kualitas yang baik. Ini sangat penting dalam membuat keputusan yang tepat tentang struktur data, karena memilih jumlah klaster yang optimal memastikan bahwa data diorganisir dengan cara yang paling masuk akal dan bermanfaat, baik untuk analisis lebih lanjut maupun untuk aplikasi praktis dalam pengelolaan data.

2.2.7 Visualisasi Data

Visualisasi data adalah proses representasi data dalam bentuk grafis atau visual, seperti grafik, diagram, peta, atau infografis [18]. Tujuan utama dari visualisasi data adalah untuk membuat data kompleks lebih mudah dipahami dan dianalisis [19]. Dengan menggunakan elemen visual seperti grafik batang, diagram lingkaran, atau peta panas, informasi yang terkandung dalam data dapat disampaikan dengan cara yang lebih intuitif dan menarik dibandingkan dengan hanya menggunakan angka atau teks. Visualisasi data memungkinkan pengguna untuk dengan cepat melihat pola, tren, dan anomali yang mungkin tersembunyi dalam data mentah.

Tujuan utama dari visualisasi data adalah untuk memperjelas informasi, memungkinkan pengambilan keputusan yang lebih baik, dan menyampaikan data dengan cara yang mudah dipahami oleh berbagai audiens [18]. Dengan visualisasi yang baik, data yang kompleks dapat diinterpretasikan dengan cepat, yang sangat penting dalam dunia bisnis, penelitian, dan berbagai bidang lainnya. Visualisasi membantu dalam mengenali hubungan antara variabel, mengidentifikasi tren historis, dan memprediksi hasil masa depan. Selain itu, visualisasi data juga memainkan peran penting dalam komunikasi, mempermudah presentasi informasi kepada stakeholder atau publik yang mungkin tidak memiliki latar belakang teknis yang kuat [19].

Salah satu alasan utama visualiasasi data ini penting untuk proses analisis adalah karena otak manusia lebih mudah memproses informasi visual daripada teks atau angka. Visualisasi data membantu dalam mengatasi keterbatasan ini dengan mengubah data menjadi bentuk yang lebih sederhana dan mudah dipahami. Ini memungkinkan lebih cepat mengidentifikasi masalah dan peluang, serta membuat keputusan yang lebih baik dan lebih cepat. Selain itu, visualisasi data juga meningkatkan keterlibatan dan retensi informasi. Grafik dan gambar sering kali lebih

menarik dan lebih mudah diingat dibandingkan dengan data tabel atau laporan teks. Oleh karena itu, dalam lingkungan bisnis yang serba cepat dan penuh dengan informasi, visualisasi data menjadi alat yang sangat penting untuk analisis data yang efektif dan komunikasi informasi yang efisien.

Ada berbagai jenis bentuk visualisasi data yang dapat digunakan tergantung pada jenis data dan informasi yang ingin disampaikan. Beberapa bentuk visualisasi yang umum meliputi grafik batang (bar chart) untuk membandingkan kategori, diagram lingkaran (pie chart) untuk menunjukkan proporsi, grafik garis (line chart) untuk mengilustrasikan tren dari waktu ke waktu, peta panas (heatmap) untuk menunjukkan intensitas data di berbagai area, dan diagram pencar (scatter plot) untuk menampilkan hubungan antara dua variabel. Selain itu, visualisasi lebih kompleks seperti diagram alur (flowchart), peta geografis, dan visualisasi jaringan juga digunakan untuk analisis yang lebih mendalam dan interaktif.

Untuk membuat visualisasi data, ada berbagai alat yang dapat digunakan, mulai dari perangkat lunak sederhana hingga platform analisis data yang lebih canggih. Beberapa alat populer untuk visualisasi data salah satunya adalah dengan menggunakan jupyter notebook yang berbasis pada pemrograman Python. Bahasa python ini yang memiliki pustaka khusus untuk visualisasi seperti Matplotlib. Karena itu python memungkinkan pengguna dari berbagai tingkat keahlian untuk membuat visualisasi data yang informatif dan menarik, membantu menyampaikan pesan dengan lebih efektif.

2.2.8 Contoh Kasus

1. Sumber Data

Untuk pengambilan sumber data, dengan menggunakan metode pengambilan data langsung ke SMA 5 Muhammadiyah Yogyakarta. Dimana data yang ambil merupakan data peminjaman buku di perpustakaan sekolah. Data ini memiliki banyak variable yaitu, nama peminjam, identitas peminjam, jenis kelamin peminjam, kelas peminjam , judul buku, kode exemplar, jenis buku, genre buku, penulis, penerbit, tanggal pinjam, tanggal kembali, dan keterangan pengembalian buku. Namun disini hanya akan menggunakan variabel identitas peminjam, jenis kelamin peminjam, kelas peminjam, judul buku, jenis buku, genre buku, penulis buku, tanggal pinjam, tanggal kembali, dan keterangan pengembalian buku.

2. Data Sampel

Pada perhitungan manual ini akan digunakan 3 cluster, dan mengambil sampel sejumlah 40 data dari data peminjaman perpustakaan , yang akan digunakan sebagai contoh perhitungan manual . Untuk kolom juga menggunakan kolom yang sama, agar memperjelas runtutan langkah demi langkah menggunakan perhitungan manual. Hasil dari perhitungan manual ini akan digunakan untuk membantu dalam analisis dan penjalanan langkah berikutnya yang menggunakan pemrograman python beserta librarynya.

Data awal sebelum diproses adalah sebelum Berikut

Tabel 2. 2 *Data Awal*

No	Identitas Peminjam	Jenis Kelamin	Kelas	Judul Buku	Jenis Buku	Genre buku	Penulis	Penerbit	Tanggal Pinjam	Pengembalian	Keterangan
1	Siswa	Laki-laki	XI MIA 1	BULAN TERE LIYE	Novel	Fantasy	Tere Liye	Gramedia Pustaka	31-Jul-18	6 AGS 18	KEMBALI
2	Siswa	Perempuan	XI MIA 1	ADA CINTA DI SMA	Novel	Romance	Haqi Achmad	Gagas Media	31-Jul-18	8 AGS 18	KEMBALI
3	Siswa	Laki-laki	X MIA 1	HUJAN TERE LIYE	Novel	Romance	Tere Liye	Gramedia Pustaka	1 AGS 18	7 AGS 18	KEMBALI
4	Guru	Perempuan	GURU	THE FALLEN	Novel	Romance	Lauren Kate	Mizan	2 AGS 18	14-Sep-18	KEMBALI
5	Siswa	Perempuan	XI MIA 1	ADORABLE MAN LEVELY LADY	Novel	Romance	Suju	Arosuka Publisher	3 AGS 18	8 AGS 18	KEMBALI
6	Siswa	Laki-laki	XI MIA 1	WHEN THE HEART CANT MOVE	Novel	Romance	Indria	Araska	3 AGS 18	6 AGS 18	KEMBALI
7	Siswa	Laki-laki	XI MIA 1	SURGA YG TAK DI RINDUKAN	Novel	Drama	Asma Nadia	Asma Nadia Jakarta	6 ags 18	27-Nov-18	KEMBALI
……											
54	Siswa	Perempuan	X IIS 1	SEASONS TO REMEMBER	Novel	Drama	Ilana Tan	Gramedia Pustaka	6 ags 18	16 AGS 18	KEMBALI
55	Siswa	Perempuan	X IIS 2	5 CM	Novel	Drama	Donny Dhirgantoro	Balai Pustaka	8 AGS 18	10 AGS 18	KEMBALI

56	Siswa	Perempuan	XI MIA 2	WHEN THE HEART CANT MOVE	Novel	Romance	Indria	Araska	9 AGS 18	15 AGS 18	KEMBALI
57	Siswa	Laki-laki	X MIA 1	BUKAN 3 IDIOT	Novel	Komedi	Boim Lebon	Indiva Media Kreasi	30 AGS 18	05-Sep-18	KEMBALI
58	Siswa	Perempuan	XII IPS 2	EKONOMI XII	Paket	Umum	Anik Widiastuti	Cempaka Putih	03-Sep-19	27-Nov-18	KEMBALI
59	Siswa	Perempuan	X MIA 1	HUJAN TERE LIYE	Novel	Romance	Tere Liye	Gramedia Pustaka	03-Sep-19	11-Nov-18	KEMBALI
60	Siswa	Perempuan	XII IPA 1	KIMIA XII	Paket	Umum	Anis Dyah Rufaida	Intan Pariwara	07-Sep-18	11-Nov-18	KEMBALI

3. Cleaning data

Proses pembersihan/cleaning yaitu membersihkan data yang variabelnya tidak memenuhi dalam perhitungan k-means. Di dapat 15 data yang bisa di pakai untuk proses perhitungan dari 18 data peminjaman buku. Untuk hasil dapat dilihat pada table 2.3.

Tabel 2. 3 *Cleaning Data*

No	Identitas Peminjam	Jenis Kelamin	Kelas	Judul Buku	Jenis Buku	Genre buku	Penulis	Penerbit	Tanggal Pinjam	Pengembalian	Keterangan
1	Siswa	Laki-laki	XI MIA 1	BULAN TERE LIYE	Novel	Fantasy	Tere Liye	Gramedia Pustaka	31-Jul-18	6 AGS 18	KEMBALI
2	Siswa	Perempuan	XI MIA 1	ADA CINTA DI SMA	Novel	Romance	Haqi Achmad	Gagas Media	31-Jul-18	8 AGS 18	KEMBALI

3	Siswa	Laki-laki	X MIA 1	HUJAN TERE LIYE	Novel	Romance	Tere Liye	Gramedia Pustaka	1 AGS 18	7 AGS 18	KEMBALI
4	Guru	Perempuan	GURU	THE FALLEN	Novel	Romance	Lauren Kate	Mizan	2 AGS 18	14-Sep-18	KEMBALI
5	Siswa	Perempuan	XI MIA 1	ADORABLE MAN LEVELY LADY	Novel	Romance	Suju	Arosuka Publisher	3 AGS 18	8 AGS 18	KEMBALI
6	Siswa	Laki-laki	XI MIA 1	WHEN THE HEART CANT MOVE	Novel	Romance	Indria	Araska	3 AGS 18	6 AGS 18	KEMBALI
7	Siswa	Laki-laki	XI MIA 1	SURGA YG TAK DI RINDUKAN	Novel	Drama	Asma Nadia	Asma Nadia Jakarta	6 ags 18	27-Nov-18	KEMBALI
……											
34	Siswa	Perempuan	X IIS 1	SEASONS TO REMEMBER	Novel	Drama	Ilana Tan	Gramedia Pustaka	6 ags 18	16 AGS 18	KEMBALI
35	Siswa	Perempuan	X IIS 2	5 CM	Novel	Drama	Donny Dhirgantoro	Balai Pustaka	8 AGS 18	10 AGS 18	KEMBALI
36	Siswa	Perempuan	XI MIA 2	WHEN THE HEART CANT MOVE	Novel	Romance	Indria	Araska	9 AGS 18	15 AGS 18	KEMBALI
37	Siswa	Laki-laki	X MIA 1	BUKAN 3 IDIOT	Novel	Komedi	Boim Lebon	Indiva Media Kreasi	30 AGS 18	05-Sep-18	KEMBALI
38	Siswa	Perempuan	XII IPS 2	EKONOMI XII	Paket	Umum	Anik Widiastuti	Cempaka Putih	03-Sep-19	27-Nov-18	KEMBALI
39	Siswa	Perempuan	X MIA 1	HUJAN TERE LIYE	Novel	Romance	Tere Liye	Gramedia Pustaka	03-Sep-19	11-Nov-18	KEMBALI
40	Siswa	Perempuan	XII IPA 1	KIMIA XII	Paket	Umum	Anis Dyah Rufaida	Intan Pariwara	07-Sep-18	11-Nov-18	KEMBALI

4. Selecting data

Proses seleksi data yaitu proses pemilihan data yang akan digunakan dalam proses data mining. Disini saya hanya menggunakan atribut identitas nama buku, genre buku, penulis buku, tanggal peminjaman. Oleh karena itu atribut yang tidak gunakan akan mengalami proses seleksi.

***Tabel 2. 4** Selecting Data*

No	Judul Buku	Genre buku	Penulis	Tanggal Pinjam
1	BULAN TERE LIYE	Fantasy	Tere Liye	31-Jul-18
2	ADA CINTA DI SMA	Romance	Haqi Achmad	31-Jul-18
3	HUJAN TERE LIYE	Romance	Tere Liye	1 AGS 18
4	THE FALLEN	Romance	Lauren Kate	2 AGS 18
5	ADORABLE MAN LEVELY LADY	Romance	Suju	3 AGS 18
6	WHEN THE HEART CANT MOVE	Romance	Indria	3 AGS 18
7	SURGA YG TAK DI RINDUKAN	Drama	Asma Nadia	6 ags 18
…….				
34	SEASONS TO REMEMBER	Drama	Ilana Tan	6 ags 18
35	5 CM	Drama	Donny Dhirgantoro	8 AGS 18
36	WHEN THE HEART CANT MOVE	Romance	Indria	9 AGS 18
37	BUKAN 3 IDIOT	Komedi	Boim Lebon	30 AGS 18
38	EKONOMI XII	Umum	Anik Widiastuti	03-Sep-19
39	HUJAN TERE LIYE	Romance	Tere Liye	03-Sep-19
40	KIMIA XII	Umum	Anis Dyah Rufaida	07-Sep-18

5. Transformasi data pertama

Dari hasil seleksi akan kemudian dilakukan proses transformasi dengan menggabungkan antara judul buku, genre, dan penulis menjadi sebuah id. Setiap data unik berdasar atribut judul buku, genre, dan penulis akan membentuk id baru. Untuk mempermudahkan dalam memahami data maka sementara dibuat kolom dengan nama keterangan yang berisikan penggabungan nilai unik dari atribut judul buku, genre, dan penulis. Untuk proses transformasi pertama dapat dilihat pada table 2.5

Tabel 2. 5 *Transformasi data pertama*

Id	Keterangan	2018	2019	2021	2022	2023
1	BULAN TERE LIYE, Fantasy, Tere Liye	1	1	0	0	0
2	ADA CINTA DI SMA, Romance, Haqi Achmad	2	2	1	0	0
3	HUJAN TERE LIYE, Romance, Tere Liye	2	2	0	0	0
4	THE FALLEN, Romance, Lauren Kate	2	0	0	0	0
5	ADORABLE MAN LEVELY LADY, Romance, Suju	2	0	0	0	0
6	WHEN THE HEART CANT MOVE, Romance, Indria	2	0	0	0	0
7	SURGA YG TAK DI RINDUKAN, Drama, Asma Nadila	1	1	0	1	0
8	MIMPI SEJUTA DOLAR, Drama, Alberthiene Endah	1	1	0	0	0
9	SEASONS TO REMEMBER, Drama, Ilana Tan	2	0	0	0	0
10	5 CM, Drama, Donny Dhirgantoro	1	1	0	1	0
11	BUKAN 3 IDIOT, Comedy, Boim Lebon	1	1	0	0	0
12	EKONOMI XII, Umum, Anik Widiastuti	3	1	0	2	1
13	KIMIA XII, Umum, Anis Dyah Rufaida	4	1	0	0	0

6. Penghapusan kolom keterangan

Kemudian transformasi kedua ada menghapus kolom keterangan dan hanya menampilkan id dan frekuensi tiap tahunnya. Hal ini dilakukan untuk mempermudah dalam melakukan pembagian data menjadi beberapa cluster. Untuk hasilnya dapat dilihat pada tabel 2.6

***Tabel 2. 6** Hasil Transformation Data*

Id	2018	2019	2021	2022	2023
1	1	1	0	0	0
2	2	2	1	0	0
3	2	2	0	0	0
4	2	0	0	0	0
5	2	0	0	0	0
6	2	0	0	0	0
7	1	1	0	1	0
8	1	1	0	0	0
9	2	0	0	0	0
10	1	1	0	1	0
11	1	1	0	0	0
12	3	1	0	2	1
13	4	1	0	0	0

7. Normalisasi data

perhitungan normalisasi manual untuk setiap tahun dilakukan dengan menghitung nilai rata-rata untuk setiap tahun kemudian menghitung standar deviasi untuk setiap tahun. Proses dimulai dengan menghitung nilai rata-rata pada tahun 2018

μ (rata rata) tahun 2018 =1+2+2+2+2+2+1+1+2+1+1+3+4/13

μ 2018 = 24/13 ≈**1.846**

Kemudian dari rata rata yang didapat , dilakukan proses perhitungan standar deviasi untuk memperoleh hasil normalisasi tiap data pada tahun 2018.

Data ke – 1 tahun 2018 $(1 - 1.846)^2 \approx 0.715$

Data ke – 2 tahun 2018 $(2 - 1.846)^2 \approx 0.024$

…….

Data ke – 12 tahun 2018 $(3 - 1.846)^2 \approx 1.332$

Data ke – 13 tahun 2018 $(4 - 1.846)^2 \approx 4.644$

Kemudian proses diulang dengan tahun selanjutnya yaitu tahun 2019. Diawali dengan mencari rata rata data

μ (rata rata) tahun 2018 = 1+2+2+0+0+0+1+1+0+1+1+1/13

μ 2018 = 10/13 ≈ **0.769**

Kemudian dilakukan proses perhitungan standar deviasi untuk memperoleh hasil normalisasi tiap data pada tahun 2019.

Data ke – 1 tahun 2018 $(1 - 0.769)^2 \approx 0.059$

Data ke – 2 tahun 2018 $(2 - 0.769)^2 \approx 1.184$

…….

Data ke – 12 tahun 2018 $(1 - 0.769)^2 \approx 0.059$

Data ke – 13 tahun 2018 $(1 - 0.769)^2 \approx 0.059$

Proses terus berlanjut sampai ditemukan hasil standar deviasi tahun 2023 yang mengakhiri proses normalisasi. Untuk hasil proses normalisasi dapat dilihat pada table 2.7.

Tabel 2. 7 *Normalisasi data*

Id	2018	2019	2021	2022	2023
1	0.715	0.059	0.006	0.095	0.006
2	0.024	1.184	0.851	0.095	0.006
3	0.024	1.184	0.006	0.095	0.006
4	0.024	0.592	0.006	0.095	0.006
5	0.024	0.592	0.006	0.095	0.006
6	0.024	0.592	0.006	0.095	0.006
7	0.715	0.059	0.006	0.314	0.006
8	0.715	0.059	0.006	0.095	0.006
9	0.024	0.592	0.006	0.095	0.006
10	0.715	0.059	0.006	0.314	0.006
11	0.715	0.059	0.006	0.095	0.006
12	1.332	0.059	0.006	2.109	0.851
13	4.644	0.059	0.006	0.095	0.006

8. Proses perhitungan *K-means*

1. Penentuan nilai k

Yang pertama dilakukan adalah menghitung nilai K, karena akan dilakukan clustering dengan nilai K yang dipakai K =3. Nilai klaster pertama untuk buku yang kurang diminati, klaster kedua untuk buku yang cukup diminati tapi menurun, klaster ketiga untuk buku yang masih dipinjam

2. Penentuan pusat cluster iterasi 1

Untuk penentuan centroid ini dengan mengambil 3 centroid dari total keseluruhan data. Centroid ini akan digunakan sebagai pusat data dan penentuan kelompok cluster tiap bagiannya. Dengan pengambilan pusat clusternya :

1. Titik No Buku 1 (0.715, 0.059, 0.006, 0.095, 0.006) dengan titik pusat m1 (0.715, 0.059, 0.006, 0.095, 0.006)

$$D(1) = \sqrt{(0.715 - 0.715)^2 + \cdots + (0.006 - 0.006)^2} = 0$$

2. Titik No Buku 1 (0.715, 0.059, 0.006, 0.095, 0.006) dengan titik pusat m2 (0.024, 1.184, 0.851, 0.095, 0.006)

$$D(2) = \sqrt{(0.715 - 0.024)^2 + \cdots + (0.006 - 0.006)^2} = 1.564$$

3. Titik No Buku 1 (0.715, 0.059, 0.006, 0.095, 0.006) dengan titik pusat m3 (1.332, 0.059, 0.006, 2.109, 0.851)

$$D(2) = \sqrt{(0.715 - 1.332)^2 + \cdots + (0.006 - 0.851)^2} = 2.269$$

3. Proses ini terus berlanjut sampai semua jarak antar cluster ditemukan. Untuk perhitungan jarak antar klaster iterasi 1 dapat dilihat pada table 2.8

Tabel 2. 8 *Jarak antar klaster*

Jarak ke m1	Jarak ke m2	jarak ke m3	Cluster terdekat
0.000	1.568	2.270	m1
1.568	0.000	2.909	m2
1.320	0.845	2.783	m2
0.873	1.032	2.601	m1
0.873	1.032	2.601	m1
0.873	1.032	2.601	m1
0.219	1.583	2.078	m1
0.000	1.568	2.270	m1
0.873	1.032	2.601	m1
0.219	1.583	2.078	m1
0.000	1.568	2.270	m1
2.270	2.909	0.000	m3
3.969	4.829	3.967	m3

Dapat diketahui kelompok data sebagai berikut :

1. Cluster 1 atau Kelompok buku kurang diminati = {1,4,5,6,7,8,9,10,11}
2. Cluster 2 atau kelompok buku cukup diminati = {2,3}
3. Cluster 3 atau kelompok buku banyak dipinjam = {13,14}

4. Perhitungan *Between Cluster Variation (BCV) dan Within Cluster Variation (WCV)*

Kemudian dilanjutkan dengan perhitungan rasio antara besaran Between Cluster Variation (BCV) dengan Within Cluster Variation (WCV) [20] , Untuk menghitung Between Cluster Variation (BCV) menggunakan persamaan 5:

$$BCV = \sum_{i,j=0}^{n} d(mi, mj) \qquad (5)$$

Dimana i dan j adalah pusat kelompok. Karena pusat kelompok ada 3 dan d(mi,mj) menyatakan jarak cluster, Maka penyelesaian BCV menjadi sebagai berikut:

1. $D(m1, m2) = \sqrt{(0.715 - 0.024)^2 + \cdots + (0.006 - 0.006)^2} = 1.564$
2. $D(m1, m3) = \sqrt{(0.715 - 1.332)^2 + \cdots + (0.006 - 0.851)^2} = 2.269$
3. $D(m2, m3) = \sqrt{(0.0240 - 1.3320)^2 + \cdots + (0.0060 - 0.8510)^2} = 2.908$

Dimana ditemukan jarak antar masing- masing cluster kemudian dapat dihitung nilai BCVnya.

BCV = 1.564 +2.269 + 2.908 = 6,741

Kemudian untuk menghitung WCV dapat dilihat pada table 2.9

Tabel 2. 9 *Perhitungan WCV*

Jarak minimum nilai claster	Kuadrat jarak (Cluster pangkat 2)
0	0
0	0
0,845	0,714
0,873	0,762
0,873	0,762
0,873	0,762
0,219	0,047
0	0
0,873	0,762
0.219	0,047
0	0
0	0
3.967	15,737
Nilai WCV	19,593

5. Penentuan rasio iterasi 2

Temukan nilai dan tentukan rasionya dengan membagi nilai BCV dengan nilai WCV. Hasil pembagian ini akan menentukan apakah iterasi akan dilanjutkan atau dihentikan. Jika rasio yang diperoleh lebih besar dari rasio sebelumnya, iterasi dilanjutkan [21]. Jika rasio tersebut kurang dari atau sama dengan rasio sebelumnya, iterasi dihentikan. Informasi lebih lanjut mengenai perbandingan ini dapat ditemukan di bawah ini:

$$Rasio(1) = \frac{BCV}{WCV} = \frac{6{,}741}{19{,}593} = 0{,}344$$

Karena ini merupakan iterasi pertama maka rasio sebelumnya adalah 0. Setelah diketahui rasio sekarang dan rasio sebelumnya, langkah selanjutnya adalah membandingkan kedua rasio. Disini diketahui bahwa nilai rasio sekarang > rasio sebelumnya, karena nilai rasio sekarang lebih besar dari rasio sebelumnya maka iterasi dilanjutkan

6. Iterasi 2

Menentukan nilai centroid baru dengan cara mencari rata rata dengan menjumlahkan semua nilai centroid sebelumnya dibagi dengan jumlah centroid tersebut. Lebih jelasnya mengenai pusat pusat centroid yang baru dapat dilihat pada table 2.10.

Tabel 2. 10 *Penentuan pusat centroid baru*

Pusat Centroid														
Centroid 1					Centroid 2					Centroid 3				
2018	2019	2021	2022	2023	2018	2019	2021	2022	2023	2018	2019	2021	2022	2023
0,715	0,059	0,006	0,095	0,006										
					0,024	1.184	0,851	0,095	0,006					
					0,024	1.184	0,006	0,095	0,006					
0,024	0,592	0,006	0,095	0,006										
0,024	0,592	0,006	0,095	0,006										
0,024	0,592	0,006	0,095	0,006										
0,715	0,059	0,006	0,314	0,006										
0,715	0,059	0,006	0,095	0,006										
0,024	0,592	0,006	0,095	0,006										
0,715	0,059	0,006	0,314	0,006										
0,715	0,059	0,006	0,095	0,006										
										1.332	0.059	0.006	2.109	0.851
										4.644	0.059	0.006	0.095	0.006
3,671	2,663	0,054	1,293	0,054	0,048	2368	0,857	0,19	0,012	5,976	0,118	0,012	2,204	0,857

7. Penentuan pusat cluster iterasi 2

Dari table 2.1 dapat ditemukan 3 pusat centroid yang baru, yakni :

Centroid 1 atau m1 = {3,671, 2,663, 0,054, 1,293, 0,054}

Centroid 2 atau m2 = {0,048, 2368, 0,857, 0,19, 0,012}

Centroid 3 atau m3 = {5,976, 0,118, 0,012, 2,204, 0,857}

8. Perhitungan jarak cluster iterasi 2

Dari pusat centroid yang baru, melakukan perhitungan dengan menggunakan Euclidean distance seperti ditahap iterasi pertama, hasil dari perhitungan dicari jarak terdekat dengan cluster. Perhitungan jarak pada iterasi kedua memiliki langkah yang sama dengan perhitungan jarak pada iterasi 1. Tabel 2.11 merupakan hasil perhitungan jarak masing masing data dengan pusat cluster

Tabel 2. 11 *Jarak antar klaster iterasi 2*

No	Jarak ke m1	Jarak ke m2	jarak ke m3	Cluster terdekat
1	2,567	3,578	4,567	m1
2	3,578	1,789	5,987	m2
3	4,566	1,789	2,5	m2
4	2,567	3,578	4,567	m1
5	2,567	3,578	4,567	m1
6	2,567	3,578	4,567	m1
7	2,132	3,334	3,114	m1
8	1,797	3,007	3,334	m1
9	2,567	3,007	2,5	m1
10	2,567	3,334	2,061552813	m1
11	1,797	3,334	3,578	m1
12	3,114	3,114	1,789	m3
13	2,132	2,5	1,789	m3

Dapat diketahui kelompok data sebagai berikut :

1. Cluster 1 atau Kelompok buku kurang diminati = {1,4,5,6,7,8,9,10,11}
2. Cluster 2 atau kelompok buku cukup diminati = {2,3}
3. Cluster 3 atau kelompok buku banyak dipinjam = {13,14}

9. Perhitungan (BCV) dan (WCV) iterasi 2

Menghitung nilai BCV sama dengan di iterasi pertama dimana i dan j adalah pusat kelompok. Karena pusat kelompok ada 3 dan d(mi,mj) menyatakan jarak cluster, Maka penyelesaian BCV menjadi sebagai berikut:

1. $D(m1, m2) = \sqrt{(3.671 - 0.048)^2 + \cdots + (0.054 - 0.012)^2} = 3{,}883$
2. $D(m1, m2) = \sqrt{3.671 - 5.976^2 + \cdots + (0.054 - 0.857)^2} = 3{,}642$
3. $D(m1, m2) = \sqrt{0.048 - 5.976^2 + \cdots + (0.012 - 0.857)^2} = 6{,}760$

Setelah ditemukannya jarak antar cluster kemudian dapat dihitung nilai BCV-nya BCV = 1,58673+ 2.54430 + 2,69258= 14,285

Menghitung nilai WCV sesuai di iterasi pertama dengan mencari jarak terkecil antara data dengan cluster kemudian dikuadratkan, dan jumlahkan semua hasil. Adapun perhitungan WCV iterasi 2 dapat dilihat pada table 2.12

Tabel 2. 12 *Perhitungan WCV iterasi 2*

Jarak minimum nilai claster	Kuadrat jarak (Cluster pangkat 2)
2,567	6,589489
1,789	3,200521
1,789	3,200521
2,567	6,589489
2,567	6,589489
2,567	6,589489
2,132	4,545424
1,797	3,229209
2,567	6,589489
2,567	6,589489
1,797	3,229209
1,789	3,200521
1,789	3,200521
Nilai WCV	63,34286

10. Penentuan rasio iterasi 2

Selanjutnya sama seperti tahapan iterasi pertama yaitu mencari nilai dan menentukan nilai rasio mencari nilai rasio dilakukan dengan cara membagi hasil nilai BCV dengan WCV kemudian hasil yang didapatkan dilakukan untuk penentuan apakah lanjut pada proses iterasi selanjutnya atau tidak, Jika nilai rasio sekarang lebih besar dari nilai rasio sebelumnya maka iterasi dilanjutkan, jika Jika nilai rasio sekarang lebih kecil atau sama dengan dari nilai rasio sebelumnya maka iterasi berhenti.

$$Rasio(2) = \frac{BCV}{WCV} = \frac{14{,}285}{63{,}34286} = 0{,}225$$

Dari perhitungan iterasi pertama dapat diketahui rasio sebelumnya adalah 0,344. Kemudian untuk hasil perhitungan rasio iterasi kedua didapatkan hasil 0,255. Dengan membandingkan antara kedua hasil, dimana rasio sekarang < rasio sebelumnya maka iterasi dihentikan sampai di iterasi kedua.

11. Perhitungan *Sillhouette score*

Langkah selanjutnya adalah pengecekan kualitas klister dengan silhouette score. Metode Silhouette (Silhouette Score) adalah metode evaluasi internal yang dengan tujuan untuk mengukur seberapa baik setiap data point berada dalam kelompoknya sendiri dan seberapa terpisah kelompok tersebut dari kelompok lainnya [22]. Metode *Silhouette Score* memberikan skor numerik untuk setiap data point dalam pengelompokan, dan skor tersebut berkisar antara -1 hingga 1. Nilai Silhouette Score menggambarkan seberapa baik data point tersebut cocok dengan kelompoknya sendiri. Untuk implementasi perhitungan dimulai dengan menghitung nilai a(i) dan b(i) . Untuk penentuan nilai a(i) dan b(i) diambil dari

perbandingan 2 jarak antar cluster yang ada pada table 2.9 , dan akan diambil jarak antar cluster 1 dan 2 sebagai a(i) dan b(i). Untuk lebih jelasnya dapat dilihat pada table 2.13 yang merupakan a(i) dan 2.12 yang merupakan b(i)

Tabel 2. 13 *Jarak antar cluster a(i)*

No	Jarak ke m1	Cluster terdekat
1	2,567	m1
2	3,578	m2
3	4,566	m2
4	2,567	m1
5	2,567	m1
6	2,567	m1
7	2,132	m1
8	1,797	m1
9	2,567	m1
10	2,567	m1
11	1,797	m1
12	3,114	m3
13	2,132	m3

Tabel 2. 14 *Jarak antar klaster b(i)*

No	Jarak ke m2	Cluster terdekat
1	3,578	m1
2	1,789	m2
3	1,789	m2
4	3,578	m1
5	3,578	m1
6	3,578	m1
7	3,334	m1
8	3,007	m1
9	3,007	m1
10	3,334	m1
11	3,334	m1
12	3,114	m3
13	2,5	m3

Setelah didapatkannya nilai a(i) dan b(i) kemudian dilakukan perhitungan silhouette dengan contoh perhitungan :

$$S(1) = \frac{3{,}578 - 2{,}567}{\max(3{,}578, 2{,}567)} = \frac{1{,}011}{3{,}578} = 0{,}2823$$

Perhitungan ini diteruskan sampai data selesai. Untuk hasil perhitungan jelasnya dapat dilihat pada table 2.15 dibawah.

Tabel 2. 15 *Perhitungan Sillhouette Score*

No	a(i)	b(i)	Perhitungan
1	2,567	3,578	0,2823
2	3,578	1,789	0.5
3	4,566	1,789	0.6037
4	2,567	3,578	0.2823
5	2,567	3,578	0.2823
6	2,567	3,578	0.2823
7	2,132	3,334	0.3386
8	1,797	3,007	0.3331
9	2,567	3,007	0.2763
10	2,567	3,334	0.2313
11	1,797	3,334	0.3769
12	3,114	3,114	0.0
13	2,132	2,5	0.2503
Total Silhouette			6,5801

Dari hasil yang didapatkan maka silhouette rata-rata :

total silhouette/jumlah data = 6,5801/13 = **0.506**

Dengan nilai siluet sebesar 0.506, hasil clustering memiliki tingkat pemisahan yang cukup baik. Di sini, nilai siluet mendekati 1, yang menunjukkan bahwa titik-titik dalam setiap cluster berada cukup jauh dari cluster lainnya dan dekat dengan pusat cluster tempat mereka berada. Ini menandakan bahwa clustering relatif homogen, dengan setiap titik memiliki kedekatan yang signifikan dengan anggota

dalam cluster yang sama dan jarak yang cukup besar dari anggota cluster lainnya. Dalam konteks aplikasi praktis, ini dapat diartikan bahwa pengelompokan berhasil memisahkan data dengan baik menjadi kelompok-kelompok yang berbeda, sehingga dapat memberikan wawasan yang berguna untuk analisis lebih lanjut atau pengambilan keputusan yang berkaitan dengan data tersebut.

12. Perhitungan *Davied Bouldin Score* dan *Calinski-Harabasz Index*

Perhitungan DBS dan CHI dimulai dengan mencari pusat centroid antar cluster. Untuk pusat centroid ini sudah dengan iterasi kedua seperti dibawah.

Centroid 1 atau m1 = {3,671, 2,663, 0,054, 1,293, 0,054}

Centroid 2 atau m2 = {0,048, 2,368, 0,857, 0,19, 0,012}

Centroid 3 atau m3 = {5,976, 0,118, 0,012, 2,204, 0,857}

Setelah itu cari matriks jarak dengan menghitung jarak antar centroid Menggunakan rumus dibawah.

1. $D(m1, m2) = \sqrt{(3.671 - 0.048)^2 + \cdots + (0.054 - 0.012)^2} = 3{,}883$
2. $D(m1, m2) = \sqrt{3.671 - 5.976^2 + \cdots + (0.054 - 0.857)^2} = 3{,}642$
3. $D(m1, m2) = \sqrt{0.048 - 5.976^2 + \cdots + (0.012 - 0.857)^2} = 6{,}760$

Setelah jarak antar cluster maka diketahui matiks jaraknya :

	M1	M2	M3
M1	0	3,883	3,642
M2	3,883	0	6,760
M3	3,642	6,760	0

Selanjutnya dapat dihitung nilai DBInya :

4. Rata-rata jarak centroid tiap cluster

   Rm1 = 3,883 + 3,642/2 = 3,762

   Rm2 = 3,883+6,76/2 = 5,321

   Rm3 = 3,642+6,76/2 = 5,201

5. Rasio Davies-Bouldin untuk Setiap Cluster:

   DBI m1 = $(5,321-5,201)^2$ = 0,0144

   DBI m2 = $(3,762-5,201)^2$ = 2,07

   DBI m3 = $(3,762-5,321)^2$ = 2,43

   DBI Keseluruhan = 0,0144 + 2,07 + 2,43 = 4,5144

6. Setelah diketahui rasio DBI masing masing cluster, kemudian cari DBI rata ratanya

   DBI = 4,5144/ 3 = **1,504**

Adapun untuk perhitungan CHInya :

1. Dispersi dalam tiap cluster

   Disp m1 = (3,671 + 2,663 + 0,054 + 1,293 + 0,054) =7,736.

   Disp m2 = (0,048 + 2,368 + 0,857 + 0,19 + 0,012) = 3,4788

   Disp m3 = (5,976 +0,118 + 0,012 +2,204+ 0,857) = 9,168

2. Dispersi antar cluster

$$DispBeetween = 5\ x\ (7{,}736^2 + 3{,}4788^2 + 9{,}168^2 = 101{,}914$$

3. Calinski Harabasz Index

$$CHI = \frac{(\frac{101{,}914}{3})}{(\frac{20{,}3828}{3})} = \frac{33{,}971}{6{,}794} = 5$$

Dengan Davies-Bouldin Index (DBI) sebesar 1,504 dan Calinski-Harabasz Index (CHI) sebesar 5, hasil evaluasi clustering menunjukkan pembagian cluster kurang optimal. Nilai DBI yang rendah menandakan pembagian klaster yang kurang optimal, tetapi masih ada beberapa permasalahan kemungkinan disebabkan oleh masalah seperti keberadaan noise dalam data atau data yang terlalu sedikit. Di sisi lain, nilai CHI yang relatif tinggi menunjukkan adanya pemisahan yang cukup jelas antara klaster-klaster yang dihasilkan, meskipun perlu diingat bahwa CHI tidaklah menjadi penentu tunggal dalam mengevaluasi kualitas clustering. Oleh karena itu, hasil dari perhitungan CHI ini menunjukkan bahwa ada ruang untuk perbaikan lebih lanjut dalam analisis clustering ini.

13. Representasi pengetahuan dan pembahasan

Berdasarkan hasil perhitungan dapat ditemukan kelompok cluster yang terbentuk . Sebelum melakukan proses presentasi pengetahuan dan pembahasan data yang berupa id ini perlu dikembalikan menjadi atribut judul buku, genre, dan penulis. Dari hal ini maka akan diketahui pola peminjaman buku. Untuk lebih jelasnya dapat dilihat pada table 2.16.

Tabel 2. 16 *Transformasi Hasil*

Id	Keterangan	Genre	Judul	Cluster
1	BULAN TERE LIYE	Fantasy	Tere Liye	m1
2	ADA CINTA DI SMA	Romance	Haqi Achmad	m2
3	HUJAN TERE LIYE	Romance	Tere Liye	m2
4	THE FALLEN	Romance	Lauren Kate	m1
5	ADORABLE MAN LEVELY LADY	Romance	Suju	m1

6	WHEN THE HEART CANT MOVE	Romance	Indria	m1
7	SURGA YG TAK DI RINDUKAN	Drama	Asma Nadila	m1
8	MIMPI SEJUTA DOLAR	Drama	Alberthiene Endah	m1
9	SEASONS TO REMEMBER	Drama	Ilana Tan	m1
10	5 CM	Drama	Donny Dhirgantoro	m1
11	BUKAN 3 IDIOT	Comedy	Boim Lebon	m1
12	EKONOMI XII	Umum	Anik Widiastuti	m3
13	KIMIA XII	Umum	Anis Dyah Rufaida	m3

Berdasarkan dari hasil yang diperoleh diberikan, dari hasil ini dapat diketahui kelompok cluster berdasarkan karakteristiknya :

1. **Cluster 1** (Kelompok Buku Kurang Diminati):

   Buku nomor 1, 4, 5 6, 7, 8, 9, 10, dan 11 termasuk dalam kelompok ini. Kelompok ini mencakup buku-buku yang memiliki tingkat peminjaman yang rendah atau kurang diminati oleh pengunjung perpustakaan. Cluster 1 ini berisikan buku dengan genre romance, drama, dan fantasy

2. **Cluster 2** (Kelompok Buku Cukup Diminati):

   Buku nomor 2 dan 3 termasuk dalam kelompok ini. Kelompok ini mencakup buku-buku yang memiliki tingkat peminjaman yang sedang atau cukup diminati oleh pengunjung perpustakaan. Cluster 2 ini berisikan buku dengan genre romance,

3. **Cluster 3** (Kelompok Buku Banyak Dipinjam):

   Tidak ada buku yang sesuai dengan nomor 13 dan 14 dalam data yang diberikan, sehingga informasi tentang buku yang terdapat dalam kelompok ini tidak dapat dipastikan.

BAB III. Metode Penelitian

3.1 Metode Pengumpulan Data

Ada proses pengumpulan data dan informasi dalam penelitian ini menggunakan beberapa metode antara lain :

1. Metode Wawancara

   Wawancara ini dilakukan dengan mengadakan sesi survei dengan pengelola perpustakaan dan penanggung jawab perpustakaan yang mengelola database peminjaman buku, menanyakan poin-poin masalah dan tujuan yang ingin dicapai, serta informasi yang diperlukan, yang kemudian dianalisis guna menyelesaikan permasalahan.

2. Studi Pustaka

   Teori pendukung diselidiki dalam penerapan data mining pada kelompok data menggunakan algoritma k-means untuk menganalisis referensi jurnal ilmiah, website , e- book dan literatur skripsi yang berhubungan dengan penelitian yang dilakukan.

3. Data Privat

   Data peminjaman buku di perpustakaan SMA 5 Muhammadiyah diambil dari database perpustakaan rentang tahun 2018 – 2023 dengan jumlah data 247. Data ini berisikan atribut nama peminjam, identitas peminjam, jenis kelamin peminjam, kelas peminjam, nama buku, kode exemplar, jenis buku, genre buku, penulis buku, penerbit buku, tanggal pinjam, tanggal pengembalian, dan keterangan pengembalian.

4. Studi Pustaka

Teori pendukung diselidiki dalam penerapan data mining pada kelompok data menggunakan algoritma k-means untuk menganalisis referensi jurnal ilmiah, buku, website, e- book dan literatur skripsi yang berhubungan dengan penelitian yang dilakukan.

3.2 Spesifikasi Kebutuhan

Dalam melakukan penelitian ini menggunakan perangkat lunak (software) dan perangkat keras (hardware) untuk membantu dalam proses penelitian. Berikut merupakan perangkat keras dan perangkat lunak yang digunakan dalam penelitian antara lain:

3.2.1 Perangkat Keras

Perangkat keras atau hardware yang digunakan untuk membantu dalam melakukan proses penelitian yaitu dengan menggunakan Laptop HP 14s- dk1xxx dengan spesifikasi sebagai berikut :

1. Prosessor Intel(R) Core(TM) i5-8300H.
2. RAM 4GB
3. LCD 15,6”
4. AMD Athlon Gold
5. SSD 500 MB

3.2.2 Perangkat Lunak

Perangkat lunak atau software yang digunakan dalam melakukan proses penelitian sebagai berikut :

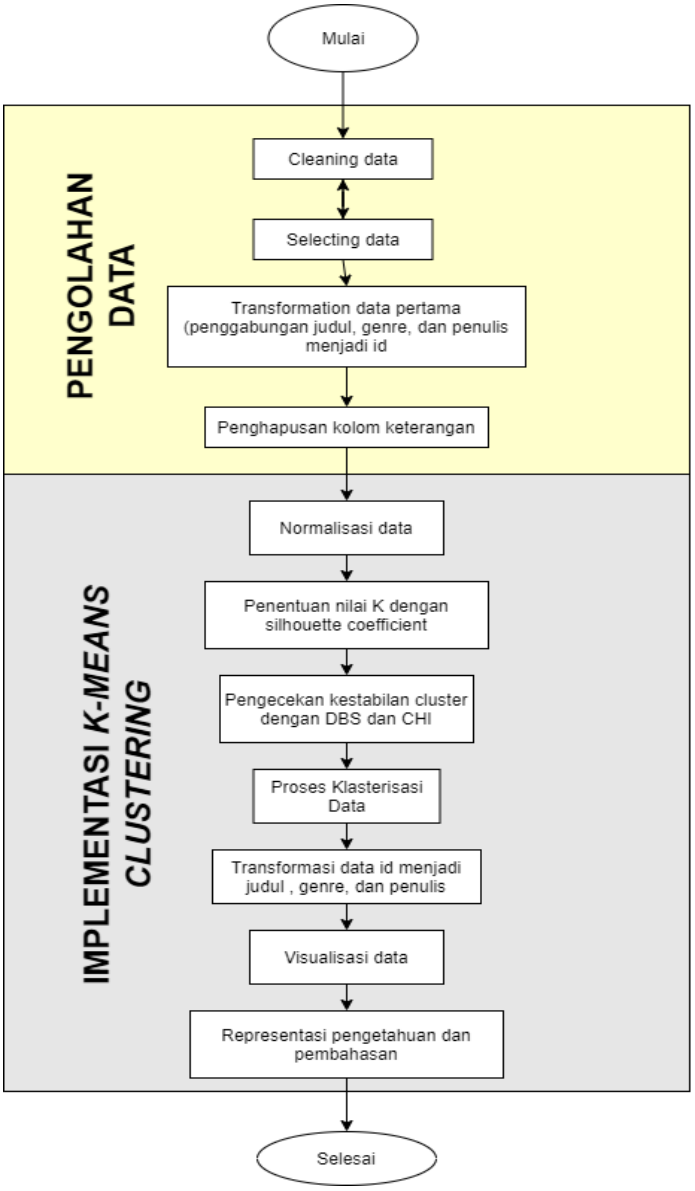
1. Sistem operasi : Windows 10 Home Single Language 64-bit
2. Microsoft Excel 2019
3. Microsoft Word 2019
4. Python v3.11.4
5. Google Chrome dan Microsoft edge
6. Anaconda promt
7. Jupyter Notebook
8. Library :
    1) from pandas preparation dataset
    2) from numpy preparation dataset
    3) from sklearn.cluster import KMeans
    4) from sklearn.metrics.pairwise import cosine_similarity
    5) from sklearn.metrics import silhouette_score

3.2.3 Kebutuhan Data

Kebutuhan data yang dibutuhkan dalam penelitian ini yaitu data yang diberikan oleh pihak perpustakaan SMA 5 Muhammadiyah dengan jumlah data sebanyak 265 dengan atribut atau variable sebanyak 2 yaitu judul buku, dan tahun peminjaman untuk mengukur skala peminjaman buku dari tahun 2018-2023.

3.2.4 Tahapan Pengolahan Data

Untuk melakukan proses pengolahan data menggunakan data mining, hal yang perlu dilakukan adalah menyiapkan data yang akan diolah, proses ini juga bisa disebut dengan data preparation. Pengolahan sendiri memiliki beberapa tahapan yang perlu dilakukan sesuai dengan flowchart 3.1 yaitu:



Flowchart 3.1 Tahapan *pengolahan data*

1. *Cleaning data*

Cleaning data adalah proses penting dalam analisis data yang melibatkan identifikasi dan penanganan data yang tidak akurat, tidak relevan, atau tidak lengkap dalam dataset [23] . Dalam data peminjaman buku perpustakaan yang memiliki jumlah 245 data dengan atribut nama peminjam, identitas peminjam, jenis kelamin, kelas, nama buku, jenis buku, genre buku, penulis buku, penerbit buku, tanggal pinjam, tanggal kembali, dan keterangan. Karena banyaknya atribut yang digunakan menyebabkan data memiliki banyak NULL, oleh karenanya perlu dilakukan proses cleaning data. Untuk mengetahui ada tidak data yang missing value atau NULL dengan menggunakan source koding.

```
df.isna().sum()
```

Adapun contoh hasil dari pengecekan dapat dilihat pada gambar 3.1 adalah sebagai berikut

```
No                    0
Nama Peminjam         1
Identitas Peminjam    1
Jenis Kelamin         1
Kelas                 1
Judul Buku            0
Kode Eksemplar        8
Jenis Buku            0
Genre buku            0
Penulis               1
Penerbit              1
Tanggal Pinjam        0
Pengembalian         70
Keterangan           61
tahun peminjaman      0
dtype: int64
```

Gambar 3. 1 *Hasil pengecekan nilai NULL*

Karena banyak data yang missing value maka perlu dilakukan proses cleaning data, proses cleaning data ini tidak boleh hanya menghapus data yang missing value saja. Pada proses cleaning data ini perlu dilakukan perkiraan data mana yang mau dihapus dan data yang dapat ditambal. Untuk menambal data bisa dengan melakukan replace menggunakan nilai rata" (apabila numerik) ataupun replace string. Pada pengujian ini karena atribut yang saya gunakan adalah data buku yang dipinjam, dan data tanggal peminjamannya, maka apabila terdapat missing value pada data buku yang dipinjam dan tanggal saya bisa drop data tersebut agar tidak mengganggu jalannya proses pengolahan data, Adapun apabila atribut lain seperti nama peminjam, identitas peminjam, jenis kelamin, kelas, tanggal kembali, dan keterangan terdapat missing value maka dapat dilakukan penambalan dengan menggunakan replace string (-) karena atribut tersebut tidak digunakan untuk pemrosesan data.

2. *Selecting data*

Seleksi data (*data selectio*n) adalah proses memilih subset atau bagian dari dataset yang relevan dan diperlukan untuk analisis atau pemodelan tertentu [24]. Dari hasil cleaning data, kemudian dilakukan lakukan proses selection data. Selection data ini biasanya digunakan untuk memilah berapa jumlah data yang ingin dipakai dan atribut apa saja yang ingin dipakai. Penelitian ini berfokus pada pengelompokan data peminjaman buku berdasarkan frekuensi pertahun, oleh karenanya dalam penelitian ini hanya menggunakan atribut judul buku, jenis buku, genre buku, penulis buku, penerbit buku, dan tanggal pinjam. Kemudian atribut akan mengalami penyeleksian berdasar kebutuhan penelitian.

3. Transformation data

Transformasi data adalah proses mengubah atau mengolah data dari bentuk asalnya ke bentuk lain, dengan tujuan untuk mempermudah analisis, mengurangi bias, atau meningkatkan performa model prediksi. Transformasi data dapat diterapkan pada data numerik maupun kategorikal, dan sering digunakan dalam analisis data dan machine learning. Karena data yang digunakan adalah data kategori dan untuk proses pengolahan data ini menggunakan *K-means Clustering*, maka data peminjaman buku perlu diubah menjadi numerik. Untuk mengubah data peminjaman buku SMA 5 Muhammadiyah yang tadinya kategori menjadi numerik, dengan menggunakan frekuensi peminjaman buku tiap tahun untuk mengetahui pola peminjaman buku berdasarkan frekuensi peminjamannya pertahun. Untuk transformasi pada penelitian ini menggunakan atribut judul buku, genre, penulis, dan tanggal pinjam. Dari atribut tersebut kemudian ditransformasikan dengan menggabungkan antara judul buku, genre, dan penulis berdasarkan keunikan data menjadi atribut id yang baru, kemudian untuk tanggal pinjam ditransformasikan menjadi tahun 2018 – 2023 dengan menampilakan frekuensinya tiap tahun.

3.2.5 Implementasi *K-Means Clustering*

1. Normalisasi Data

Normalisasi data adalah teknik statistik yang digunakan dalam pemrosesan awal data untuk membuat variabel yang berbeda menjadi lebih dapat akurat untuk dibandingkan satu sama lain. Ini seperti menerjemahkan semua “bahasa” data yang berbeda ini ke dalam satu dialek universal. Proses ini melibatkan penyesuaian skala nilai sehingga memiliki distribusi yang seragam, biasanya dengan mengatur nilai rata-rata menjadi nol dan standar deviasi menjadi satu. Dengan normalisasi, data dari berbagai

skala dan unit pengukuran dapat dibandingkan secara langsung, yang memudahkan dalam melakukan analisis lebih lanjut.

Langkah pertama dalam melakukan normalisasi data adalah mengumpulkan dan memahami dataset yang akan dinormalisasi. Ini termasuk mengidentifikasi variabel-variabel yang ada dalam dataset dan memastikan data tersebut bersih dari nilai-nilai yang hilang atau outlier yang ekstrem. Setelah itu, setiap variabel dihitung nilai rata-ratanya (mean) dan standar deviasinya. Proses ini dapat dilakukan dengan menggunakan alat statistik atau perangkat lunak pemrograman seperti Python atau R, yang memiliki fungsi bawaan untuk menghitung mean dan standar deviasi dengan mudah.

Langkah berikutnya adalah menggunakan nilai rata-rata dan standar deviasi tersebut untuk menormaliasai setiap nilai dalam variabel. Formula yang digunakan adalah mengurangi nilai rata-rata dari setiap data poin dan kemudian membaginya dengan standar deviasi. Proses ini menghasilkan nilai-nilai yang memiliki rata-rata nol dan standar deviasi satu, yang berarti data telah ternormalisasi. Setelah normalisasi, data siap untuk dianalisis lebih lanjut atau digunakan dalam model pembelajaran mesin. Normalisasi ini memastikan bahwa setiap variabel memiliki kontribusi yang setara dalam analisis dan membantu meningkatkan kinerja algoritma yang digunakan dalam analisis data.

2. *Silhouette Coeficient*

Silhouette Coefficient (Koefisien Silhouette) adalah sebuah metrik evaluasi yang digunakan untuk mengukur sejauh mana kualitas pengelompokan data hasil dari algoritma clustering, seperti *K-Means*. Metrik ini memberikan nilai antara -1 hingga 1, di mana nilai semakin mendekati 1 menandakan bahwa pengelompokan data semakin baik, sedangkan nilai mendekati -1 menandakan bahwa data lebih cocok untuk dikelompokkan di cluster lain, dan nilai sekitar 0 menandakan bahwa data berada dekat

dengan batas antara dua cluster [22]. Dalam penelitian pengelompokan peminjaman buku perpustakaan SMA 5 Muhammadiyah menggunakan metode *Silhouette Coefficient* beberapa manfaat yang dapat membantu penelitan yaitu :

1. Evaluasi Kualitas Pengelompokan: *Silhouette Coefficient* membantu dalam mengevaluasi sejauh mana kualitas pengelompokan buku-buku ke dalam cluster-cluster tertentu tanpa bergantung pada metode Elbow. Hasil evaluasi ini memberikan gambaran tentang seberapa baik cluster-cluster yang terbentuk memisahkan buku-buku berdasarkan tingkat popularitasnya. Meskipun jumlah cluster telah ditentukan sebelumnya, *Silhouette Coefficient* akan membantu memahami sejauh mana pengelompokan tersebut berhasil dalam membentuk kelompok-kelompok yang kohesif dan terpisah dengan baik.
2. Menilai Homogenitas dan Tumpang Tindih: *Silhouette Coefficient* memberikan informasi tentang sejauh mana buku-buku dalam satu cluster secara homogen, serta sejauh mana *cluster-cluster* tersebut saling tumpang tindih. Hal ini membantu dalam memahami apakah ada kelompok-kelompok yang mungkin terlalu tumpang tindih atau memiliki kelompok-kelompok dengan tingkat kesamaan yang rendah.
3. Interpretasi Hasil yang Lebih Mendalam: *Silhouette Coefficient* memberikan pemahaman lebih mendalam tentang bagaimana buku-buku dikelompokkan berdasarkan popularitasnya, sehingga hasil clustering tidak hanya berupa kelompok-kelompok yang terbentuk, tetapi juga memberikan penilaian kualitatif tentang sejauh mana kelompok-kelompok tersebut efektif dalam mengidentifikasi kategori buku. Dengan menggunakan Silhouette Coefficient, Anda dapat menilai dan menginterpretasikan hasil clustering dengan lebih baik tanpa bergantung pada jumlah cluster yang ditentukan oleh metode Elbow.

4. Evaluasi Sensitivitas Terhadap Jumlah Cluster: Dalam beberapa kasus, metode Elbow mungkin tidak selalu memberikan jumlah cluster yang optimal atau jelas. Dalam situasi ini, Silhouette Coefficient dapat membantu dalam mengevaluasi sensitivitas hasil clustering terhadap perubahan jumlah cluster, sehingga memberikan wawasan tambahan tentang variasi hasil. Dengan demikian, penggunaan Silhouette Coefficient dalam penelitian ini tetap memberikan perspektif analisis yang bermanfaat dalam mengukur kualitas pengelompokan dan memberikan interpretasi hasil clustering yang lebih mendalam tanpa harus tergantung pada metode Elbow untuk menentukan jumlah cluster.

3. Davies-Bouldin Index dan Calinski-Harabasz

Davies-Bouldin Index dan Calinski-Harabasz Index adalah dua metrik evaluasi yang sering digunakan dalam penelitian untuk mengukur kualitas pengelompokan data hasil dari algoritma *clustering*, seperti *K-Means*. Berikut adalah bagaimana kedua metrik ini dapat diterapkan dalam konteks penelitian pengelompokan peminjaman buku perpustakaan SMA 5 Muhammadiyah.

1. Evaluasi kualitas pengelompokan dengan *Davies-Bouldin Index: Davies-Bouldin Index* membantu dalam mengevaluasi seberapa baik klaster yang dihasilkan memisahkan buku-buku berdasarkan tingkat popularitasnya. Skor ini dihitung berdasarkan rasio antara jarak dalam klaster (*intra-cluster distance*) dan jarak antar klaster *(inter-cluster distance*). Semakin kecil nilai *Davies-Bouldin Index*, semakin baik klaster yang dihasilkan. Dalam penelitian ini, setelah menentukan jumlah klaster dengan metode seperti *Elbow* atau heuristik lainnya, *Davies-Bouldin Index* akan digunakan untuk menilai sejauh mana klaster yang terbentuk memiliki kepadatan yang tinggi dan terpisah dengan baik dari klaster lainnya. Hal ini memberikan gambaran tentang efektivitas pengelompokan buku-buku

tersebut dan membantu dalam memahami apakah buku-buku dalam satu klaster cukup homogen dan terpisah dengan jelas dari buku-buku di klaster lain.

2. Menilai homogenitas dan pemisahan dengan *Calinski-Harabasz Index: Calinski-Harabasz Index*, juga dikenal sebagai *Variance Ratio Criterion*, digunakan untuk mengukur seberapa baik klaster-klaster tersebut tersebar dan memiliki variabilitas rendah dalam klaster. Indeks ini dihitung berdasarkan rasio antara jumlah dispersi antar klaster dan jumlah dispersi dalam klaster. Semakin tinggi nilai Calinski-Harabasz Index, semakin baik klaster yang dihasilkan. Dalam konteks penelitian ini, setelah klaster terbentuk, *Calinski-Harabasz Index* akan digunakan untuk menilai homogenitas buku-buku dalam klaster dan seberapa baik klaster-klaster tersebut terpisah. Ini membantu dalam mengidentifikasi apakah ada klaster yang mungkin terlalu luas atau terlalu sempit, serta memastikan bahwa pengelompokan yang dilakukan tidak hanya terfokus pada jumlah klaster tetapi juga pada kualitas klaster yang dihasilkan.
3. Interpretasi hasil yang lebih mendalam: Dengan menggunakan Davies-Bouldin Index dan Calinski-Harabasz Index, penelitian ini dapat memberikan pemahaman yang lebih mendalam tentang efektivitas pengelompokan buku berdasarkan popularitasnya. Kedua metrik ini memberikan penilaian kuantitatif yang jelas tentang kualitas klaster yang dihasilkan, sehingga hasil clustering tidak hanya berupa kelompok-kelompok yang terbentuk tetapi juga memberikan wawasan tentang seberapa baik kelompok-kelompok tersebut mengidentifikasi kategori buku yang ada. Evaluasi ini membantu dalam memahami apakah buku-buku dalam klaster tertentu benar-benar memiliki kesamaan yang tinggi dan apakah klaster-klaster tersebut terpisah dengan jelas, memberikan keyakinan bahwa pengelompokan yang dilakukan efektif dan berarti.

4. Evaluasi sensitivitas terhadap jumlah klaster: Dalam beberapa kasus, metode seperti Elbow mungkin tidak selalu memberikan jumlah klaster yang optimal atau jelas. Dalam situasi ini, Davies-Bouldin Index dan Calinski-Harabasz Index dapat membantu dalam mengevaluasi sensitivitas hasil clustering terhadap perubahan jumlah klaster. Dengan menghitung nilai-nilai ini untuk berbagai jumlah klaster, peneliti dapat menilai apakah perubahan jumlah klaster memberikan peningkatan signifikan dalam kualitas klaster yang dihasilkan. Ini memberikan wawasan tambahan tentang variasi hasil dan membantu dalam membuat keputusan yang lebih tepat tentang struktur data yang dihasilkan oleh proses klasterisasi, memastikan bahwa jumlah klaster yang dipilih benar-benar optimal dan sesuai untuk analisis lebih lanjut.

4. Kmeans Clustering

K-Means Clustering adalah teknik analisis data yang mengelompokan objek data menjadi beberapa kelompok atau cluster berdasarkan kesamaan karakteristiknya. Dua jenis grup data yang biasa digunakan untuk mengelompokan data, yaitu grup data hierarkis dan grup data non-hierarkis. K-Means adalah metode pengelompokan data non-hierarkis yang mencoba membagi data yang ada menjadi satu atau lebih cluster/cluster [9]. Berdasarkan permasalahan yang dialami selama wawancara , yaitu berkurangnya minat baca buku di SMA 5 Muhammadiyah dikarenakan adanya pandemi covid 19. Penelitian ini bertujuan untuk menemukan pola peminjaman buku berdasarkan frekuensi peminjamannya tiap tahun dengan membaginya menjadi 3 cluster yaitu buku paling diminati, cukup diminati, dan kurang diminati berdasarkan karakteristik tiap clusternya.

5. Visualisasi data

Visualisasi data adalah langkah penting dalam penelitian untuk membantu memahami dan mengomunikasikan hasil analisis secara efektif. Dalam konteks penelitian pengelompokan peminjaman buku perpustakaan SMA 5 Muhammadiyah, visualisasi dengan menggunakan grafik bar membantu untuk dengan jelas menampilkan informasi tentang tren dan pola peminjaman buku berdasarkan klaster yang dihasilkan dari proses klasterisasi. Setelah proses klasterisasi dilakukan, langkah berikutnya adalah menghitung nilai maksimum jumlah peminjaman buku per tahun untuk setiap klaster. Hal ini dilakukan untuk mendapatkan gambaran tentang buku-buku mana yang paling sering dipinjam dalam setiap klaster dan bagaimana tren peminjaman berubah dari tahun ke tahun. Nilai maksimum ini memberikan informasi yang lebih jelas dan tidak tumpang tindih, dibandingkan dengan menggunakan nilai rata-rata yang bisa saja menghasilkan nilai yang kurang jelas akibat adanya variasi yang besar dalam data. Dengan menggunakan matplotlib, grafik bar dibuat untuk memvisualisasikan nilai maksimum peminjaman buku per tahun untuk setiap klaster. Berikut adalah langkah-langkah dalam proses visualisasi ini:

1. **Menyiapkan Data untuk Grafik**: Data yang telah dikelompokkan dan dihitung nilai maksimumnya diatur dalam format yang sesuai untuk pembuatan grafik bar. Data ini mencakup tahun-tahun peminjaman sebagai sumbu x dan nilai maksimum peminjaman sebagai sumbu y.
2. **Membuat Grafik Bar**: Grafik bar dibuat dengan masing-masing bar mewakili nilai maksimum peminjaman buku per tahun untuk setiap klaster. Lebar bar ditentukan sedemikian rupa agar setiap klaster memiliki satu set bar yang terpisah tetapi berdekatan untuk memudahkan perbandingan visual.

3. **Menambahkan Label dan Penanda**: Setiap bar diberi label yang menunjukkan nilai maksimum peminjaman untuk memberikan informasi yang lebih mendetail kepada penonton grafik. Penanda ini membantu dalam memahami dan membandingkan nilai-nilai antar klaster dengan lebih mudah.
4. **Menyesuaikan Tampilan Grafik**: Aspek visual lainnya seperti judul grafik, label sumbu, gridlines, dan legenda ditambahkan untuk meningkatkan keterbacaan dan interpretasi grafik. Judul grafik memberikan gambaran umum tentang apa yang ditampilkan, sementara label sumbu dan legenda membantu dalam memahami skala dan kategori yang ditampilkan.

Analisis dan Interpretasi Visualisasi: Setelah grafik bar dibuat, langkah selanjutnya adalah menganalisis dan menginterpretasikan hasil visualisasi. Peneliti melihat pola dan tren yang muncul dari grafik untuk memahami bagaimana peminjaman buku berubah dari tahun ke tahun dalam setiap klaster. Analisis ini membantu dalam menarik kesimpulan tentang minat pengunjung perpustakaan terhadap berbagai kategori buku dan bagaimana buku-buku tersebut dapat dikelompokkan secara efektif untuk mendukung manajemen perpustakaan yang lebih baik. Dengan mengikuti metodologi ini, penelitian dapat menghasilkan visualisasi yang jelas dan informatif, membantu dalam memahami tren peminjaman buku dan mendukung pengambilan keputusan yang lebih baik dalam manajemen perpustakaan.

BAB IV. Hasil Dan Pembahasan

1.1 Hasil Pengumpulan Data

Merujuk pada data peminjamanan buku perpustakaan SMA 5 Muhammadiyah mulai dari tahun 2018 sampai tahun 2023 yang berisikan atribut nama peminjam, identitas peminjam, jenis kelamin peminjam, kelas peminjam, nama buku, kode exemplar, jenis buku, genre buku, penulis buku, penerbit buku, tanggal pinjam, tanggal pengembalian, dan keterangan pengembalian. Kemudian data tersebut diolah dalam penelitian untuk mengetahui kelompok cluster beradasarkan karakterisitik masing – masing, dan menjadi bahan untuk membantu mempermudah pihak perpusyakaan dalam melakukan analisis guna pengadaan buku.

1.2 Pengolahan Data

Pengembangan sistem dilakukan dengan cara mengolah data peminjaman buku odi Perpustakaan SMA 5 Muhammadiyah menggunakan metode data mining k-means clustering untuk proses klasterisasi data, metode silhouette coefficient untuk pengujian kualitas jumlah cluster, dan metode David Bouldin Score dan Calinski Harabasz Index untuk pengujian kerapatan dan kestabilan cluster. pengembangan sistem ini akan dilakukan sesuai dengan tahapan data mining. Proses pengolahan data ini akan dilakukan dengan 2 tahap yaitu preprocessing data , dan implementasi metode K-Means Clustering

1.2.1 Data Awal

Pada tabel 4.1 merupakan data awal yang didapat dari perpustakaan SMA 5 Muhammadiyah , data awal merupakan peminjaman buku di SMA 5 Muhammadiyah dari tahun 2018 sampai tahun 2023 yang memiliki jumlah 246 data peminjaman buku. Berikut merupakan sebagian tampilan dari data peminjaman buku tersebut.

***Tabel 4. 1** Data Awal*

No	Nama Peminjam	Identitas Peminjam	Jenis Kelamin	Kelas	Judul Buku	Kode Eksemplar	Jenis Buku	Genre buku	Penulis	Penerbit	Tanggal Pinjam	Pengembalian	Keterangan	tahun peminjaman
1	Denata	Siswa	Laki-laki	Xi Mia 1	Bulan Tere Liye	0788.SMA	Novel	Fantasy	Tere Liye	Gramedia Pustaka	31/07/2018	06/08/2018	Kembali	2018
2	Inayah	Siswa	Perempuan	Xi Mia 1	Ada Cinta Di Sma	KODE KOSONG	Novel	Romance	Haqi Achmad	Gagas Media	31/07/2018	08/08/2018	Kembali	2018
3	Raeyhan	Siswa	Laki-laki	X Mia 1	Hujan Tere Liye	0737.SMA	Novel	Romance	Tere Liye	Gramedia Pustaka	01/08/2018	07/08/2018	Kembali	2018
4	Evi Widiastui	Guru	Perempuan	Guru	The Fallen	0367.SMA	Novel	Romance	Lauren Kate	Mizan	02/08/2018	14-Sep-18	Kembali	2018
5	Inayah	Siswa	Perempuan	Xi Mia 1	Adorable Man Levely Lady	0546.SMA	Novel	Romance	Suju	Arosuka Publisher	03/08/2018	08/08/2018	Kembali	2018
6	Denata	Siswa	Laki-laki	Xi Mia 1	When The Heart Cant Move	KODE KOSONG	Novel	Romance	Indria	Araska	03/08/2018	06/08/2018	Kembali	2018
7	Denata	Siswa	Laki-laki	Xi Mia 1	Surga Yg Tak Di Rindukan	KODE KOSONG	Novel	Drama	Asma Nadia	Asma Nadia Jakarta	06/08/2018	27-Nov-18	Kembali	2018
8	Isna	Siswa	Perempuan	X Iis 1	Mimpi Sejuta Dolar	KODE KOSONG	Novel	Drama	Alberthiene Endah	Gramedia Pustaka Utama	06/08/2018	16/08/2018	Kembali	2018
9	Alya	Siswa	Perempuan	X Iis 1	Seasons To Remember	0797.SMA	Novel	Drama	Ilana Tan	Gramedia Pustaka	06/08/2018	16/08/2018	Kembali	2018

10	Yeti	Siswa	Perempuan	Xii Ips 3	Matematika Xii Sem 1 2004	17256.H	Paket	Umum	Abdur Rohman	Kementrian pendidikan	07/08/2018		Kembali	2018
11	Siti Nur	Siswa	Perempuan	X Iis 2	5 Cm	0018.SMA	Novel	Drama	Donny Dhirgantoro	Balai Pustaka	08/08/2018	10/08/2018	Kembali	2018
12	Rifan	Siswa	Laki-laki	Xi Mia 2	Si Dul Anak Jakarta	14937.SMA	Novel	Drama	Aman Datuk Madjoindo	Balai Pustaka	09/08/2018			2018
13	Adrina	Siswa	Perempuan	Xi Mia 2	When The Heart Cant Move	KODE KOSONG	Novel	Romance	Indria	Araska	09/08/2018	15/08/2018	Kembali	2018
14	Raeyhan	Siswa	Laki-laki	X Mia 1	Bukan 3 Idiot	,0414.SMA	Novel	Komedi	Boim Lebon	Indiva Media Kreasi	30/08/2018	05-Sep-18	Kembali	2018
15	Nadianti	Siswa	Perempuan	Xii Ips 2	Ekonomi Xii	19406.SMA	Novel	Umum	Anik Widiastuti	Cempaka Putih	03/09/2019	27-Nov-18	Kembali	2019
16	Apriwlan	Siswa	Perempuan	X Mia 1	Hujan Tere Liye	0737.SMA	Novel	Romance	Tere Liye	Gramedia Pustaka	03/09/2019	11-Nov-18	Kembali	2019
17	Raeyhan	Siswa	Laki-laki	X Mia 1	Radikus Makankakus	0416.SMA	Novel	Komedi	Raditya Dika	GagasMedia	05/09/2018			2018
18	Mayla Noor	Siswa	Perempuan	Xii Ipa 1	Kimia Xii	KODE KOSONG	Paket	Umum	Anis Dyah Rufaida	Intan Pariwara	07/09/2018	11-Nov-18	Kembali	2018
19	Yeti	Siswa	Perempuan	Xii Ips 3	Personality Plus	0284.SMA	Novel	Eduksi	Littauer	Karisma Publishing Group	10/09/2018	25-Sep-18	Kembali	2018

1.2.2 Preproccessing data

Preprocessing data adalah serangkaian langkah yang dilakukan pada data mentah sebelum data tersebut digunakan dalam analisis data [25]. Tujuannya adalah untuk membersihkan, mentransformasi, dan mempersiapkan data agar sesuai dengan kebutuhan analisis atau model yang akan digunakan. Untuk melakukan tahapan *preproccessing* ini menggunakan aplikasi anaconda promt dan bahasa *python*. Adapun tahapan *preproccesing data* adalah sebagai berikut.

1. *Import library*

   Library python adalah kumpulan modul terkait berisi kumpulan kode yang dapat digunakan berulang kali dalam program yang berbeda. Adanya *library* membuat pemrograman *python* menjadi lebih sederhana dan nyaman bagi *programmer* karena tidak perlu menulis kode yang sama berulang kali untuk program yang berbeda [26] .Python adalah bahasa pemrograman yang paling cocok digunakan untuk pengolahan data karena memiliki berbagai library dan framework yang kuat untuk pengolahan data seperti NumPy, pandas, scikit-learn,dan matplotlib. Library ini menawarkan berbagai fungsi dan alat yang diperlukan untuk membantu dalam analisis, manipulasi, dan pemodelan data dengan mudah dan efisien. Pada Kode 4.1 merupakan kode untuk memanggil *library python* yang akan digunakan untuk proses pengolahan data ini. Untuk library yang digunakan meliputi library pandas yang akan digunakan untuk membantu membaca, mengubah, dan memanipulasi data, *library numpy* yang digunakan untuk perhitungan, *library matplotlib* untuk proses visualisasi data, dan *library skylearn* untuk processing data yaitu untuk membantu penentuan cluster, pembakuan nilai *data standar scaler* , perhitungan *sillhouette score, davies bouldin score, dan Calinski Harabasz Index*

```
1. import pandas as pd
2. import matplotlib.pyplot as plt
3. import numpy as np
4. from sklearn.cluster import KMeans
5. from sklearn.preprocessing import StandardScaler
6. from sklearn.metrics import silhouette_score
7. from sklearn.datasets import make_blobs
8. from sklearn.metrics import
   davies_bouldin_score, calinski_harabasz_score
```

Kode 4. 1 *Library Python*

2. Pembacaan Data

Pembacaan data dalam konteks pemrograman merujuk pada proses mengambil atau memuat data dari sumber eksternal (seperti file, basis data, atau API) ke dalam struktur data yang bisa digunakan dalam program. Dalam Python, khususnya dengan menggunakan pustaka pandas, pembacaan data sering dilakukan dari berbagai format file seperti CSV, Excel, JSON, dan lainnya. Karena data menggunakan format excel, maka untuk kode pembacaan data dapat dilihat pada kode 4.2 dibawah [27].

```
1. df = pd.read_excel('data.xlsx')
2. df
```

Kode 4. 2 *Reading Data*

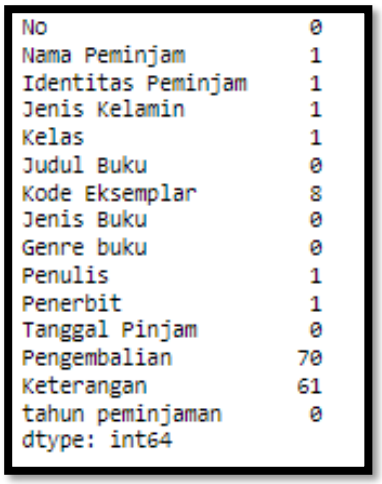
3. Cleaning data

Proses *cleaning data* adalah proses memperbaiki atau menghapus data yang tidak konsisten. Tahapan ini diperlukan sebelum analisis data karena biasanya data awal mengandung informasi yang tidak akurat, tidak tersusun rapi, atau tidak lengkap. Sebelum dilakukan proses cleaning data, perlu di check terlebih dahulu apakah data masih memiliki nilai yang tidak konsisten atau tidak. Pada kode 4.3 merupakan kode yang digunakan untuk melakukan pengecekan nilai *NaN* atau *NULL*.

```
1.  df.isna().sum()
```

Kode 4. 3 *Checking Data*

Untuk hasil dari pengecekannya dapat dilihat pada gambar 4.1 dibawah ini.



Gambar 4. 1 *Hasil Pengecekan Data*

Berdasarkan gambar 4.1 yang merupakan gambar hasil pengecekan diatas, dapat diketahui banyak terdapat nilai *NULL* pada data. Nilai *NULL* ini terdapat pada atribut yang memiliki type data string. Oleh karenanya perlu dilakukan cleaning data dengan menggunakan metode *formatting cell* untuk mengatasi ketidakkonsistenan data. Kode 4.3 menampilkan proses cleaning data dengan menggunakan metode *formatting cell*. Caranya adalah mengisi nilai yang tidak konsisten ini dengan simbol, kata, atau kalimat yang diinginkan.

```
1. df = df.fillna('-')
2.df['tahun peminjaman'] = df['tahun peminjaman'].astype(int)
Df
```

Kode 4.4 *Cleaning data*

Untuk hasil *cleaning data* menggunakan kode 4.3 ini dapat dilihat pada table 4.2. Dapat dilihat bahwa sudah tidak terdapat nilai yang tidak konsisten karena sudah digantikan dengan simbol “-“.

Tabel 4. 2 *Hasil Cleaning Data*

No	Nama Peminjam	Identitas Peminjam	Jenis Kelamin	Kelas	Judul Buku	Kode Eksemplar	Jenis Buku	Genre buku	Penulis	Penerbit	Tanggal Pinjam	Pengembalian	Keterangan	tahun peminjaman
1	Denata	Siswa	Laki-laki	Xi Mia 1	Bulan Tere Liye	0788.SMA	Novel	Fantasy	Tere Liye	Gramedia Pustaka	31/07/2018	06/08/2018	Kembali	2018
2	Inayah	Siswa	Perempuan	Xi Mia 1	Ada Cinta Di Sma	KODE KOSONG	Novel	Romance	Haqi Achmad	Gagas Media	31/07/2018	08/08/2018	Kembali	2018
3	Raeyhan	Siswa	Laki-laki	X Mia 1	Hujan Tere Liye	0737.SMA	Novel	Romance	Tere Liye	Gramedia Pustaka	01/08/2018	07/08/2018	Kembali	2018
4	Evi Widiastui	Guru	Perempuan	Guru	The Fallen	0367.SMA	Novel	Romance	Lauren Kate	Mizan	02/08/2018	14-Sep-18	Kembali	2018
5	Inayah	Siswa	Perempuan	Xi Mia 1	Adorable Man Levely Lady	0546.SMA	Novel	Romance	Suju	Arosuka Publisher	03/08/2018	08/08/2018	Kembali	2018
6	Denata	Siswa	Laki-laki	Xi Mia 1	When The Heart Cant Move	KODE KOSONG	Novel	Romance	Indria	Araska	03/08/2018	06/08/2018	Kembali	2018
7	Denata	Siswa	Laki-laki	Xi Mia 1	Surga Yg Tak Di Rindukan	KODE KOSONG	Novel	Drama	Asma Nadia	Asma Nadia Jakarta	06/08/2018	27-Nov-18	Kembali	2018
8	Isna	Siswa	Perempuan	X Iis 1	Mimpi Sejuta Dolar	KODE KOSONG	Novel	Drama	Alberthiene Endah	Gramedia Pustaka Utama	06/08/2018	16/08/2018	Kembali	2018
9	Alya	Siswa	Perempuan	X Iis 1	Seasons To Remember	0797.SMA	Novel	Drama	Ilana Tan	Gramedia Pustaka	06/08/2018	16/08/2018	Kembali	2018
10	Yeti	Siswa	Perempuan	Xii Ips 3	Matematika Xii Sem 1 2004	17256.H	Paket	Umum	Abdur Rohman	Kementrian	07/08/2018	-	Kembali	2018

										pendidikan				
……														
238	Auria Nada	Siswa	Perempuan	Xii Ips	Sosiologi Xii	19192	Paket	Umum	Poerwanti Hadi Pratiwi	Kementrian Pendidikan	04/03/2023	-	Kembali	2023
239	Auria Nada	Siswa	Perempuan	Xii Ips	Ekonomi Xii	04742	Paket	Umum	Prapto Muntoko, M.Pd	Kementrian Pendidikan	04/03/2023	-	Kembali	2023
240	Mezallunа	Siswa	Perempuan	Xii Ips	Surga Yg Tak Di Rindukan	KODE KOSONG	Novel	Drama	Asma Nadia	Asma Nadia Jakarta	04/03/2018	-	Kembali	2018
241	Mezallunа	Siswa	Perempuan	Xii Ips	Bahasa Indonesia Xii	Kode Eksemplar	Paket	Umum	Maman Suryaman, Dr, M.Pd	Kementrian Pendidikan	04/03/2022	07-Mar-22	Kembali	2022
242	Adinda	Siswa	Perempuan	Xii Ips	Bahasa Indonesia Xii	Kode Eksemplar	Paket	Umum	Maman Suryaman, Dr, M.Pd	Kementrian Pendidikan	04/03/2022	24-Mar-22	Kembali	2022
243	Era Fanigera	Siswa	Perempuan	Xii Ips	Sosiologi Xii	04772	Paket	Umum	Poerwanti Hadi Pratiwi	Kementrian Pendidikan	15/03/2022	-	-	2022
244	Era Fanigera	Siswa	Perempuan	Xii Ips	Detik Sosiologi	Kode Eksemplar	Paket	Umum	Poerwanti Hadi Pratiwi	Kementrian Pendidikan	15/03/2022	-	-	2022
245	Tri Rahayu	PPL UAD	Perempuan	Luar	Menjelajah Dunia Biologi 1	02495.SMA	Paket	Umum	Endah Sulistyowati	Kementrian Pendidikan	12/05/2023	26/05/2023	-	2023

4. Selecting data

Setelah dilakukan cleaning data, kemudian proses dilanjutkan dengan *selecting data*. *Selecting data* adalah proses memilih atribut tertentu dari awal berdasarkan kriteria tertentu. Kode 4.5 merupakan function untuk melakukan selecting atribut data menggunakan *function iloc*. Untuk lebih jelasnya dapat dilihat pada kode 4.5 dibawah ini.

```
1. df = df.iloc[:, [5,8,9,14]]
2. df
```

***Kode 4. 5** Selecting data*

Dapat dilihat pada kode 4.5 , terjadi selecting data dengan memilih atribut ke 5,7,8,9,10, dan 14. Atribut ini dihitung berdasarkan nilai indexnya, dimana index akan diawali dengan 0 sampai jumlah atribut -1. Atirbut yang digunakan disini adalah nama peminjam, judul buku, jenis buku, genre buku, penulis, dan tanggal peminjaman. Untuk hasil dari selecting data ini dapat dilihat pada tabel 4.3

***Tabel 4. 3** Selecting Data*

No	Judul Buku	Genre buku	Penulis	tahun peminjaman
1	Bulan Tere Liye	Fantasy	Tere Liye	2018
2	Ada Cinta Di Sma	Romance	Haqi Achmad	2018
3	Hujan Tere Liye	Romance	Tere Liye	2018
4	The Fallen	Romance	Lauren Kate	2018
5	Adorable Man Levely Lady	Romance	Suju	2018
6	When The Heart Cant Move	Romance	Indria	2018
7	Surga Yg Tak Di Rindukan	Drama	Asma Nadia	2018
8	Mimpi Sejuta Dolar	Drama	Alberthiene Endah	2018
9	Seasons To Remember	Drama	Ilana Tan	2018
10	Matematika Xii Sem 1 2004	Umum	Abdur Rohman	2018
.....				

238	Sosiologi Xii	Umum	Poerwanti Hadi Prati wi	2023
239	Ekonomi Xii	Umum	Prapto Muntoko, M.Pd	2023
240	Surga Yg Tak Di Rindukan	Drama	Asma Nadia	2018
241	Bahasa Indonesia Xii	Umum	Maman Suryaman, Dr, M.Pd	2022
242	Bahasa Indonesia Xii	Umum	Maman Suryaman, Dr, M.Pd	2022
243	Sosiologi Xii	Umum	Poerwanti Hadi Prati wi	2022
244	Detik Sosiologi	Umum	Poerwanti Hadi Prati wi	2022
245	Menjelajah Dunia Biologi 1	Umum	Endah Sulistyowati	2023
238	Sosiologi Xii	Umum	Poerwanti Hadi Prati wi	2023

5. Transformasi Data

Transformasi data merupakan proses mengubah bentuk data menjadi bentuk yang akan memudahkan untuk tahap pemrosesan dan analisis data. Transformasi data ini sangat luas bentuk, karena data memiliki kebutuhan yang berbeda beda untuk dilakukan analisis. Untuk melakukan transformasi data ini menggunakan function yang ada di pandas, seperti yang dapat dilihat pada kode 4.6 dibawah ini.

```
1. df['ID'] = df['Judul Buku'] + '_' + df['Genre buku'] +
   '_' + df['Penulis']
2. df = df.groupby(['ID', 'tahun
   peminjaman']).size().unstack(fill_value=0)
3. df.columns.name = None
4. df = df.reset_index()
5. df.insert(0, 'id', range(1, len(df) + 1))
6. df.rename(columns={'ID': 'keterangan'}, inplace=True)
7. df
8. nama_file_excel = "Keterangan ID.xlsx"
9. df.to_excel(nama_file_excel, index=False)
```

Kode 4. 6 *Trasnformasi Data*

Pada kode 4.6 terjadi proses transformasi data pada dataFrame berisi informasi tentang buku-buku yang dipinjam dari perpustakaan. Pertama, kolom 'Judul Buku', 'Genre

Buku', dan 'Penulis Buku' digabungkan menjadi satu ID dengan menggunakan operator '+' untuk kemudian dijadikan sebagai indeks baru DataFrame. Langkah berikutnya adalah mengelompokan data berdasarkan ID dan tahun peminjaman, dan menghitung frekuensi kemunculan setiap kombinasi ID dan tahun menggunakan fungsi groupby dan size, lalu mengubah format DataFrame agar ID menjadi kolom biasa.

Kolom ID kemudian diubah menjadi nomor urut dimulai dari 1 dengan menggunakan fungsi insert. Nama kolom 'ID' kemudian diubah menjadi 'keterangan' dengan menggunakan fungsi rename. DataFrame yang telah diolah selanjutnya disimpan dalam format Excel dengan menyertakan nama file yang telah ditentukan sebelumnya, tanpa menyertakan indeks DataFrame. Dengan demikian, kodingan ini memungkinkan untuk menghasilkan data xlsx baru dengan nama Keterangan ID yang berisi informasi terkait buku-buku yang dipinjam beserta keterangan tambahan berupa nomor urut dan frekuensi peminjaman, yang dapat digunakan untuk analisis lebih pada proses selanjutnya. Gambar 4.2 merupakan hasil dari proses transformasi menggunakan kode tersebut

	id	keterangan	2018	2019	2021	2022	2023
0	1	5 CM_Drama_Donny Dhirgantoro	1	1	1	0	0
1	2	9 SUMMERS 10 AUTUMNS_Drama_Iwan Setyawan	1	0	0	0	0
2	3	A COFFE TIME DIARY_Romance_Riri Ansar	1	0	0	0	0
3	4	ADA CINTA DI SMA_Romance_Haqi Achmad	4	3	0	2	0
4	5	ADORABLE MAN LEVELY LADY_Edukasi_Rere	1	0	0	0	0
...	...	...	...	...	...	...	...
122	123	TOMODACHI SCHOOL_Drama_Winna Efendi	0	1	0	0	0
123	124	UN SMA KING USBN 2019_Umum_FORUM TENTOR INDONESIA	0	0	0	1	0
124	125	WE GOT MARRIED_Romance_SK	1	2	0	0	0
125	126	WHEN THE HEART CANT MOVE_Romance_Indria	2	0	0	0	0
126	127	YOUTH ADAGIO_Drama_Alberta Natasia Adji	0	1	0	0	0

127 rows × 7 columns

Gambar 4. 2 *Transformasi data sementara(awal)*

Dapat dilihat bahwa terjadi transformasi data dengan mengkategorikan judul buku, genre, dan penulis menjadi id yang dimulai dari 1. Data tranformasi sementara ini akan di simpan untuk digunakan pada pengekstrakan hasil pengelompokan cluster agar dapat diketahui kelompok judul, genre, dan penulis buku berdasarkan cluster yang didapatkannya. Oleh karena itu perlu dilakukan proses transformasi lagi untuk dilakukannya proses klasterisasi. Lebih jelasnya dapat dilihat pada kode 4.7 dibawah ini.

```
1. df = pd.read_excel("Keterangan ID.xlsx")
2. df.drop(columns=['keterangan'], inplace=True)
3. df
4. nama_file_excel = "Hasil transformasi.xlsx"
5. df.to_excel(nama_file_excel, index=False)
```

Kode 4. 7 *Transformasi Data*

Kode 4.7 dimulai dengan dibaca dari file "Keterangan ID.xlsx" yang telah disimpan sebelumnya. Langkah kedua menghapus kolom 'keterangan' dari DataFrame, yang mungkin tidak relevan untuk analisis selanjutnya. Setelah itu, DataFrame yang telah dimodifikasi ditampilkan untuk memverifikasi perubahan yang telah dilakukan. Berikutnya, hasil tersebut disimpan dalam file Excel baru dengan nama "Hasil transformasi.xlsx". Untuk Hasil transformasinya dapat dilihat pada table 4.4

Tabel 4. 4 *Transformasi Data*

Id	2018	2019	2021	2022	2023
1	1	1	1	0	0
2	1	0	0	0	0
3	1	0	0	0	0
4	4	3	0	2	0
5	1	0	0	0	0
6	1	0	0	0	0
7	5	0	0	0	0
8	0	0	1	0	0
9	0	0	1	0	0
10	0	0	1	0	0

11	0	1	0	0	0
12	14	7	0	0	0
13	9	6	0	0	0
14	0	1	0	0	0
15	0	1	0	0	0
16	1	0	0	0	0
17	0	1	0	0	0
18	3	0	0	3	0
19	0	0	0	0	1
20	1	0	0	0	0
21	0	1	0	0	0
22	0	1	0	0	0
23	0	1	1	0	0
24	0	0	0	1	0
25	1	2	0	1	0
26	0	1	0	0	0
27	0	1	0	0	0
28	2	1	0	0	0
29	0	1	0	0	0
30	1	1	0	0	0
31	1	0	0	0	0
32	0	1	0	0	0
33	0	1	0	0	0
34	2	0	0	0	0
35	0	1	0	0	0
36	2	0	0	0	0
37	1	0	0	0	0
38	1	0	0	0	0
39	0	1	0	0	0
40	0	0	0	1	0
41	1	0	0	0	0
42	0	1	0	0	0
43	1	0	0	0	0
44	1	0	0	0	0
45	0	1	0	0	0
46	1	1	0	1	1
47	0	0	1	0	0
48	0	1	0	0	0
49	1	0	0	0	0
50	0	1	0	0	0
51	0	1	0	0	0
52	3	1	0	0	0
53	0	1	0	0	0

54	0	0	0	1	0
55	0	1	0	0	0
56	0	1	0	0	0
57	1	1	0	0	0
58	2	1	0	0	0
59	0	2	0	0	0
60	0	1	0	0	0
61	1	0	0	0	0
62	0	1	0	0	0
63	0	0	0	1	0
64	1	0	0	0	0
65	1	0	0	0	0
66	1	0	0	0	0
67	0	0	0	1	0
68	0	0	0	1	0
69	0	1	0	0	0
70	0	1	0	0	0
71	1	0	0	0	0
72	1	0	0	0	0
73	1	0	0	0	0
74	0	0	0	1	0
75	1	0	0	0	0
76	1	0	0	0	0
77	1	0	0	0	0
78	1	0	0	0	0
79	1	0	0	0	0
80	0	0	0	0	1
81	0	1	0	0	0
82	1	1	0	0	0
83	0	1	0	0	0
84	0	2	0	0	0
85	0	0	0	0	1
86	0	1	0	0	0
87	0	1	0	0	0
88	1	0	0	0	0
89	1	0	0	0	0
90	0	1	0	0	0
91	1	0	0	0	0
92	0	1	0	0	0
93	0	1	0	0	0
94	0	1	0	1	0
95	1	0	0	0	0
96	0	1	0	0	0

97	0	1	0	0	0
98	0	1	0	0	0
99	1	1	0	0	0
100	1	0	0	0	0
101	1	0	0	0	0
102	0	1	0	0	0
103	0	1	0	0	0
104	0	0	0	0	1
105	1	0	0	0	0
106	0	0	0	1	0
107	0	0	0	1	0
108	1	2	0	0	0
109	0	1	0	0	0
110	1	0	0	0	0
111	0	1	0	1	1
112	2	1	2	0	0
113	0	1	0	0	0
114	0	1	0	0	0
115	0	0	1	0	0
116	17	8	3	2	1
117	1	0	0	0	0
118	1	0	0	0	0
119	2	0	0	0	0
120	0	1	0	0	0
121	0	1	0	0	0
122	0	1	0	0	0
123	0	1	0	0	0
124	0	0	0	1	0
125	1	2	0	0	0
126	2	0	0	0	0
127	0	1	0	0	0

1.3 Tahap Implementasi *K-Means Clustering*

Setelah dilakukan tahap preprocessing , data sudah dianggap layak untuk digunakan. Pada proses selanjutnya adalah proses perhitungan Kmeans Clustering. Untuk tahap ini merupakan implementasi dari konsep kecerdasan buatan dan data mining menggunakan bahasa pemrograman python.

1.3.1 Normalisasi data

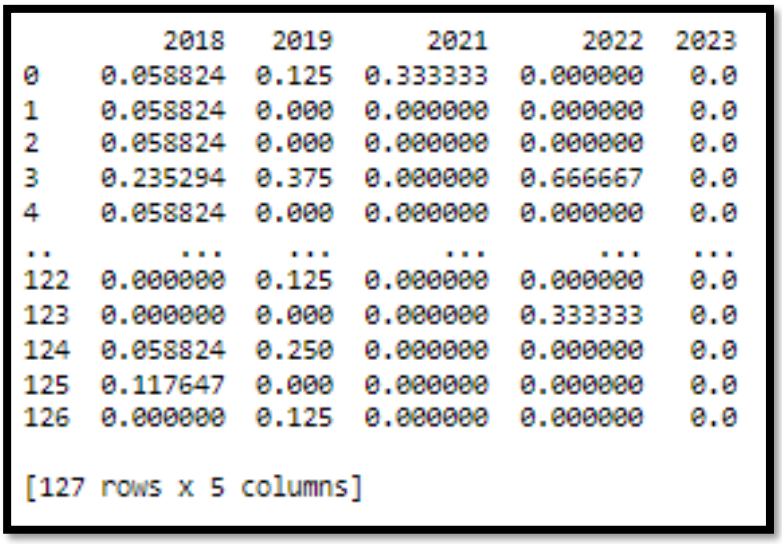
Normalisasi data adalah langkah penting dalam preprocessing data untuk analisis dan machine learning. Proses ini memastikan bahwa fitur data memiliki skala yang seragam, yang dapat meningkatkan kinerja dan efisiensi algoritma analisis. Selain itu dengan adanya normalisasi data menjadikan data menjadi lebih akurat sehingga memudahkan dalam proses penentuan jumlah cluster yang optimal menggunakan silhouette coefficient, dan evaluasi kestabilan cluster menggunakan *Calinski-Harabasz Index dan Davies-Bouldin Score.* Proses ini menggunakan library seperti pandas dan scikit-learn, menjadikan normaliasasi data dapat dilakukan dengan mudah dan efisien [28]. Untuk proses normalisasi data menggunakan python dapat dilihat pada kode 4.8 dibawah ini.

```
1. data_without_title = data.drop(columns=['Judul Buku'])
2. scaler = MinMaxScaler()
3. data_normalized = scaler.fit_transform(data_without_title)
4. data_normalized_df = pd.DataFrame(data_normalized,
   columns=data_without_title.columns)
5. print(data_normalized_df)
```

Kode 4. 8 *Normalisasi Data*

Seperti yang dapat dilihat pada kode 4.8 ada 5 fungsi yang digunakan untuk melakukan proses normalisasi data. fungsi pertama digunakan untuk Menghapus kolom 'Judul Buku' dari DataFrame karena kolom ini mungkin berisi data non-numerik yang tidak perlu dinormalisasi. Fungsi kedua untuk membuat objek MinMaxScaler untuk melakukan normalisasi data ke dalam rentang [0, 1]. Fungsi ketiga Menerapkan Min-Max Scaling pada data yang sudah dihapus kolom 'Judul Buku' dan menghasilkan data berupa array numpy dengan nilai yang dinormalisasi. Fungsi keempat digunakan untuk Mengonversi array numpy hasil normalisasi kembali menjadi DataFrame dan memberikan nama kolom yang sesuai dengan DataFrame asli tanpa kolom 'Judul Buku' yang sudah di drop. Terakhir fungsi kelima yang digunakan untuk menampilkan DataFrame

yang sudah dinormalisasi. Untuk hasil normalisasi dapat dilihat pada gambar 4.2 dibawah



	2018	2019	2021	2022	2023
0	0.058824	0.125	0.333333	0.000000	0.0
1	0.058824	0.000	0.000000	0.000000	0.0
2	0.058824	0.000	0.000000	0.000000	0.0
3	0.235294	0.375	0.000000	0.666667	0.0
4	0.058824	0.000	0.000000	0.000000	0.0
..	...	...	...	...	...
122	0.000000	0.125	0.000000	0.000000	0.0
123	0.000000	0.000	0.000000	0.333333	0.0
124	0.058824	0.250	0.000000	0.000000	0.0
125	0.117647	0.000	0.000000	0.000000	0.0
126	0.000000	0.125	0.000000	0.000000	0.0

[127 rows x 5 columns]

***Gambar 4. 3** Normalisasii Data*

Hasil dari normalisasi pada gambar 4.3 ini berupa array numpy. Nilai tersebut memiliki rentang 0 sampai 1 karena menggunakan *function min max scaler* untuk mengatur rentang data. Data hasil normalisasi ini juga mempertahankan hubungan relatif antara nilai-nilai asli. Dapat dilihat

1.3.2 Pemilihan jumlah cluster menggunakan *Sillhouette Coefficient*

Penentuan jumlah cluster adalah langkah krusial dalam analisis clustering yang sangat mempengaruhi kualitas dan interpretabilitas hasil clustering. Menentukan jumlah cluster yang optimal membantu dalam mengungkap struktur data yang mendasarinya dan memastikan bahwa cluster yang dihasilkan relevan dan bermakna. Jumlah cluster yang terlalu sedikit dapat menyebabkan informasi penting hilang karena beberapa kelompok data digabung menjadi satu, sedangkan jumlah cluster yang terlalu banyak dapat membuat hasil sulit diinterpretasikan dan bisa mengarah pada *overfitting*. Oleh karena itu, jumlah cluster yang tepat sangat penting untuk memastikan bahwa cluster yang dihasilkan kompak (objek-objek dalam cluster yang

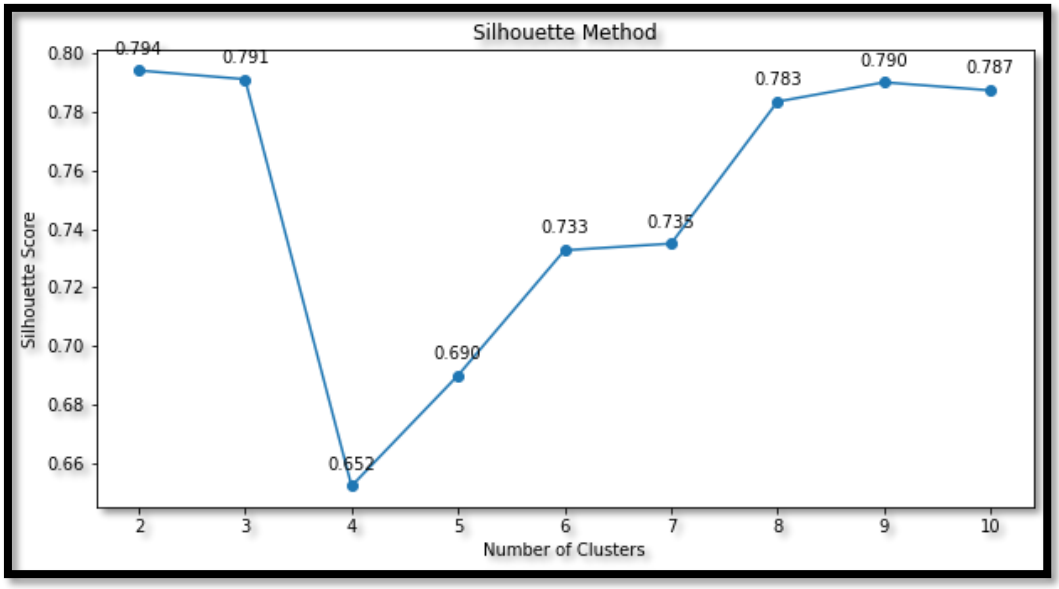
sama mirip satu sama lain) dan terpisah dengan baik (objek-objek dalam cluster yang berbeda tidak mirip). Jumlah cluster pada penelitian ini akan menggunakan 3 cluster yang akan dijadikan cluster buku paling diminati, cukup diminati, dan kurang diminati. Untuk mengecek kualitas 3 cluster ini diawali dengan menggunakan *silhouette coefficient*.

Silhouette Score adalah salah satu metode yang sering digunakan untuk menentukan jumlah cluster yang optimal. Skor ini mengukur seberapa mirip objek dengan cluster mereka sendiri dibandingkan dengan klaster lain. Nilai *Silhouette Score* berkisar antara -1 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa objek lebih dekat ke cluster mereka sendiri dibandingkan dengan cluster lain. Oleh karena itu, *Silhouette Score* dapat digunakan untuk menilai kualitas *clustering* dengan berbagai jumlah cluster dan membantu dalam memilih jumlah cluster yang menghasilkan nilai *Silhouette Score* tertinggi, yang menunjukkan *clustering* dengan kompaksi dan pemisahan yang baik [17]. Dengan demikian, penggunaan *Silhouette Score* dalam penentuan jumlah cluster dapat memastikan bahwa hasil clustering tidak hanya bermakna tetapi juga berkualitas tinggi . Untuk grafik *silhouette score* dapat dilihat dari Kode 4.8 dibawah ini. menyajikan grafik *silhouette score* dengan rentang jumlah cluster 2 sampai 10 yang bisa digunakan untuk penentuan jumlah klaster.

```
1. num_clusters = range(2, 11)
2. silhouette_scores = []
3. for k in num_clusters:
4.     kmeans = KMeans(n_clusters=k)
5.     kmeans.fit(data_normalized_df)
6.   silhouette_scores.append( silhouette_score(
   data_normalized_df , kmeans.labels_))
7.  plt.figure(figsize=(10, 5))
8.  plt.plot(num_clusters, silhouette_scores, marker='o')
9.  plt.xlabel('Number of Clusters')
10. plt.ylabel('Silhouette Score')
11. plt.title('Silhouette Method')
12. plt.show()
```

Kode 4. 9 *Grafik Sillhouette Score*

Kode tersebut menyajikan grafik *silhouette score* dengan rentang jumlah cluster 2 sampai 10 yang bisa digunakan untuk penentuan jumlah cluster pada fungsi pertama. Kemudian mendeklarasikan silhouette score sebagai array untuk menampung hasil perhitungan silhouette scorenya pada fungsi kedua. Setelah itu menggunakan perulangan untuk perhitungan silhouette score tiap cluster pada fungsi nomer 3 sampai 6. Kemudian setelah nilai silhouette score semua berhasil didapat, divisualisasikan menggunakan grafik plotpoin tiap titik jumlah silhouette scorenya. Adapun hasil dari kode 4.8 dapat dilihat pada gambar



Gambar 4. 4 *Silhouette Coefficient*

Dari Dari grafik Silhouette Score yang dihasilkan, nilai tertinggi mendekati saat jumlah cluster adalah 2. Meskipun nilai Silhouette Score optimal tercapai dengan 2 cluster, nilai untuk 3 cluster juga cukup tinggi dan memadai untuk analisis. Jika Menggunakan , jika menggunakan hanya 2 cluster, kelompok yang terbentuk terlalu umum dan kurang informatif, tidak mencerminkan variasi yang sebenarnya dalam data. Sebaliknya jika menggunakan 9 cluster sebenenarnya bisa memberikan interpretasi yang lebih detail karena memiliki score yang cukup tinggi di peringkat ketiga, tetapi untuk pengelolaan data di perpustakaan, terlalu banyak cluster dapat menjadi rumit untuk dianalisis oleh pihak perpustakaan. Selain membingungkan perpustakaan dalam menganalisis karakteristik, juga dapat membingungkan pihak perpustakaan dalam menentukan pengadaan buku berdasarkan kebutuhan mereka.

Oleh karenanya Dengan menggunakan 3 cluster, hasil clustering tetap memiliki kualitas yang baik, lebih informatif, dan lebih mudah untuk diimplementasikan dan dipahami oleh tim perpustakaan. Pendekatan ini memungkinkan pemetaan yang efektif terhadap koleksi buku, memudahkan pengelompokan dan manajemen data, serta mendukung pengambilan keputusan yang lebih efisien dan tepat.

Oleh karena itu, menetapkan jumlah cluster menjadi 3 adalah pilihan yang tepat. Alasan utamanya adalah agar dapat mengelompokan buku menjadi kategori kurang diminati, cukup diminati, dan paling diminati, sehingga tidak ada potensi buku yang cukup diminati terabaikan dengan bergabung ke dalam kelompok yang tidak diminati hanya karena menggunakan 2 cluster.

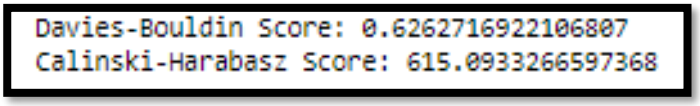
1.3.3 Pengecekan kestabilan cluster dengan CHI dan DBS

Pengukuran kestabilan jumlah cluster sangat penting untuk memastikan bahwa hasil clustering yang diperoleh adalah optimal dan dapat diandalkan. Dua metrik yang sering digunakan untuk mengevaluasi kualitas clustering adalah *Calinski-Harabasz Index dan Davies-Bouldin Score* . Kedua metrik ini memberikan perspektif yang berbeda tetapi saling melengkapi tentang kualitas clustering. CHI menekankan pada pemisahan antara cluster dan kompaksi dalam cluster dengan menggunakan rasio variabilitas, sedangkan DBS fokus pada rasio jarak antar-cluster dan kompaksi dalam-cluster. Menggunakan kedua metrik ini bersama-sama memberikan pandangan yang lebih holistik tentang kualitas clustering. Misalnya, jika jumlah cluster yang diuji memiliki nilai CHI yang tinggi dan nilai DBS yang rendah, ini mengindikasikan bahwa jumlah cluster tersebut optimal. Untuk penelitian ini karena sudah ditentukan akan menggunakan jumlah 3 kluster, maka akan dilakukan pengecekan pada kestabilan jumlah cluster ini menggunakan CHI dan DBS. Kode 4.9 menampilkan implementasi CHS dan DBS menggunakan python.

```
1. X, _ = make_blobs(n_samples=300, centers=4,
   cluster_std=0.60, random_state=0)
2. data_normalized_df = X
3. kmeans = KMeans(n_clusters=3)
4. kmeans.fit(data_normalized_df)
5. db_score = davies_bouldin_score(data_normalized_df,
   kmeans.labels_)
6. ch_score =
   calinski_harabasz_score(data_normalized_df,
   kmeans.labels_)
7. print(f"Davies-Bouldin Score: {db_score}")
8. print(f"Calinski-Harabasz Score: {ch_score}")
9. plt.figure(figsize=(12, 6))
10.      plt.subplot(1, 2, 1)
11.      plt.title('Davies-Bouldin Score')
12.      plt.scatter(X[:, 0], X[:, 1],
   c=kmeans.labels_, cmap='viridis', alpha=0.5,
   edgecolor='k')
13.      plt.xlabel('Feature 1'
14.      plt.ylabel('Feature 2')
15.      plt.colorbar()
16.      plt.subplot(1, 2, 2)
17.      plt.title('Calinski-Harabasz Score')
18.      plt.scatter(X[:, 0], X[:, 1],
   c=kmeans.labels_, cmap='viridis', alpha=0.5,
   edgecolor='k')
19.      plt.xlabel('Feature 1')
20.      plt.ylabel('Feature 2')
21.      plt.colorbar()
22.      plt.tight_layout()
23.      plt.show()
```

Kode 4. 10 *Calinski-Harabasz Score dan Davies-Bouldin Score*

Untuk hasil dari kode 4.19 dapat dilhat pada gambar 4.5 dibawah yang menampilkan hasil pemisahan data menggunakan CHI dan DBSnya menggunakan pemrograman python.



Gambar 4. 5 *Hasil Perhitungan CHS dan DBS*

Dari hasil perhitungan untuk 3 kluster, dengan *Davies-Bouldin Score* sebesar 0.626 dan *Calinski-Harabasz Index* sebesar 615,093 , dapat disimpulkan bahwa clustering tersebut menunjukkan kualitas kerapatan dan kestabilan cluster yang baik. Nilai *Davies-Bouldin index* yang rendah menandakan bahwa klaster-klaster memiliki kompaksi yang baik dan terpisah dengan jelas, sementara *Calinski-Harabasz index* yang tinggi dengan nilai diatas 100 menunjukkan bahwa cluster-cluster tersebut cukup baik dalam memisahkan data. Oleh karena itu, berdasarkan metrics silhouette, CHS, dan DBS dapat diketahui digunakannya 3 cluster untuk analisis dinilai sesuai dan dapat dianggap stabil dan optimal dalam konteks pengelolaan data. Karena hasil perhitungan dinilai sudah cukup optimal maka dapat dilanjutkan ke proses berikutnya.

1.3.4 Proses klasterisasi data

Setelah melakukan penentuan cluster dan memperhitungkan kestabilan jumlah cluster terbu, kemudian data dapat dilakukan proses klasterisasi dengan membagi menjadi 3 cluster. Adapun Untuk kode pengklasterannya dapat dilihat pada kode 4.11 dibawah ini

```
1. km = KMeans(n_clusters=3)
2. y_predicted = km.fit_predict(df.iloc[:, 1:])
3. df['cluster'] = y_predicted
4. print(df)
```

Kode 4. 11 *Klasterisasi Data*

Pada langkah pertama, model KMeans diinisialisasi dengan 3 kluster untuk melakukan proses klasterisasi. Langkah berikutnya adalah melakukan prediksi kluster untuk setiap sampel

dalam dataset menggunakan model yang telah diinisialisasi, yang menghasilkan label klaster untuk setiap sampel. Kemudian, hasil prediksi kluster ini ditambahkan sebagai kolom baru bernama 'cluster' ke dalam DataFrame awal, memungkinkan untuk melihat hasil klasterisasi untuk setiap entri dalam dataset. Dengan demikian, DataFrame yang diperbarui dengan tambahan kolom kluster memberikan gambaran tentang bagaimana setiap entri atau sampel dalam dataset dikelompokkan ke dalam kluster yang berbeda sesuai dengan pola atau karakteristik yang dimilikinya. Adapun untuk pembagian cluster berdasarkan karakteristiknya adalah sebagai berikut :

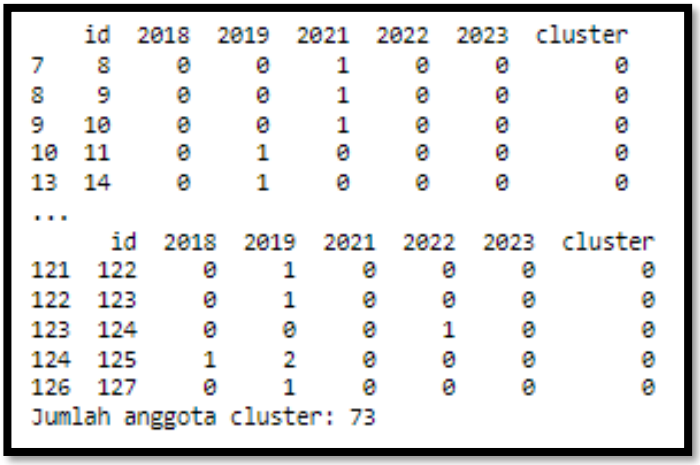
1. Cluster 1

   Karena anggota cluster 1 cukup banyak maka untuk pemanggilan cluster menggunakan kode 4.12 seperti berikut

```
1. first_cluster = df.loc[df['cluster'] == 0]
2. print(first_cluster.head(5))
3. print("...")
4. print(first_cluster.tail(5))
5. print("Jumlah anggota cluster:",
   first_cluster.shape[0])
```

Kode 4. 12 *Pemanggilan Cluster 1*

Pada kode 4.11 ini menyingkat anggota cluster 1 yang cukup banyak, caranya adalah dengan menampilkan 5 cluster diatas dan 5 cluster dibawah. Agar diketahui jumlah cluster tetap dicetak. Adapun hasilnya dapat dilihat pada gambar 4.6 dibawah ini.



	id	2018	2019	2021	2022	2023	cluster
7	8	0	0	1	0	0	0
8	9	0	0	1	0	0	0
9	10	0	0	1	0	0	0
10	11	0	1	0	0	0	0
13	14	0	1	0	0	0	0
...							
	id	2018	2019	2021	2022	2023	cluster
121	122	0	1	0	0	0	0
122	123	0	1	0	0	0	0
123	124	0	0	0	1	0	0
124	125	1	2	0	0	0	0
126	127	0	1	0	0	0	0

Jumlah anggota cluster: 73

Gambar 4. 6 *Anggota Cluster 1*

2. Cluster 2

Untuk cluster 2 yang memiliki hanya sedikit anggota maka menggunakan kode yang berbeda dan tidak perlu menampilkan jumlah anggotanya karena sudah dapat dihitung secara langsung. Adapun kode untuk memanggil cluster 2 dapat dilihat pada kode 4.13 dibawah

```
1. second_cluster = df.loc[df['cluster'] == 1]
2. second_cluster
```

Kode 4. 13 *Pemanggilan Cluster 2*

Untuk cluster kedua ini memiliki lebih sedikit anggota disbanding cluster lain. Namun meskipun demikian tetap cluster ini tetap mempertahankan karakteristiknya dari cluster lain. Adapun untuk kelompok datanya dapat dilihat pada gambar 4.7 dibawah.

	id	2018	2019	2021	2022	2023	cluster
11	12	14	7	0	0	0	1
12	13	9	6	0	0	0	1
115	116	17	8	3	2	1	1

***Gambar 4. 7** Anggota Cluster 2*

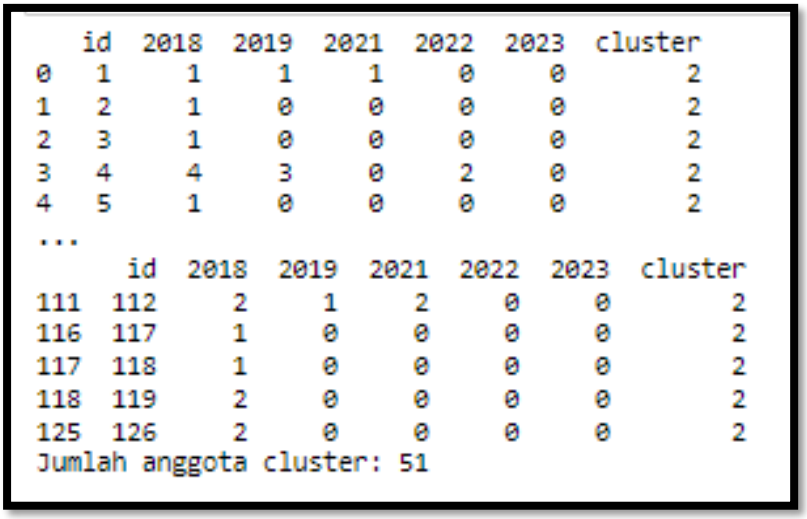
3. Cluster 3

Untuk cluster 3 memiliki jumlah anggota yang cukup banyak sama seperti cluster pertama, karena itu kode hamper menyerupai pemanggilan cluster pertama hanya saja ada sedikit perubahan . Untuk lebih jelasnya dapat melihat kode 4.14

```
1. third_cluster = df.loc[df['cluster'] == 2]
2. print(third_cluster.head(5))
3. print("...")
4. print(third_cluster.tail(5))
5. print("Jumlah anggota cluster:", third_cluster.shape[0])
```

***Kode 4. 14** Pemanggilan Cluster 3*

Hampir mirip dengan kode 4.12, Dimana kode 4.14 ini menyingkat anggota cluster 3 yang cukup banyak, caranya adalah dengan menampilkan 5 cluster diatas dan 5 cluster dibawah. Hanya saja ada perubahan deklarasi cluster menggunakan index 2 atau 3 . Agar diketahui jumlah cluster tetap dicetak. Adapun hasilnya dapat dilihat pada gambar 4.8 dibawah.



	id	2018	2019	2021	2022	2023	cluster
0	1	1	1	1	0	0	2
1	2	1	0	0	0	0	2
2	3	1	0	0	0	0	2
3	4	4	3	0	2	0	2
4	5	1	0	0	0	0	2
...							
	id	2018	2019	2021	2022	2023	cluster
111	112	2	1	2	0	0	2
116	117	1	0	0	0	0	2
117	118	1	0	0	0	0	2
118	119	2	0	0	0	0	2
125	126	2	0	0	0	0	2

Jumlah anggota cluster: 51

Gambar 4. 8 *Anggota Cluster 3*

Setelah melakukan klasterisasi menggunakan algoritma KMeans, dapat dilihat hasil pengelompokan data dibagi ke dalam 3 kluster yang berbeda. Untuk mengetahui letak pusat dari setiap kluster yang telah dibentuk, dengan menggunakan atribut "km.cluster_centers". Pusat kluster adalah titik rata-rata dari semua data dalam kluster tersebut dan mewakili posisi sentral dalam ruang fitur. Dengan melihat pusat kluster, dapat dipahami karakteristik umum dari masing-masing kluster. Adapun kode untuk menampilkan pusat 6 cluster ini dapat dilihat pada kode 4.15 dibawah ini.

```
1. print(km.cluster_centers_)
```

Kode 4. 15 *Menampilkan Pusat Cluster*

Sebelumnya dilakukan proses normalisasi pada data yang membuat data bertransformasi menjadi array numpy, oleh karenanya data akan ditampilkan dalam belum bilangan decimal.Untuk lebih jelasnya dapat dilihat pada gambar 4.9 dibawah ini.

```
[[5.47945205e-02 8.08219178e-01 8.21917808e-02 1.91780822e-01
  8.21917808e-02]
 [1.33333333e+01 7.00000000e+00 1.00000000e+00 6.66666667e-01
  3.33333333e-01]
 [1.35294118e+00 2.35294118e-01 5.88235294e-02 9.80392157e-02
  5.55111512e-17]]
```

Gambar 4. 9 *Pusat Cluster*

1.3.5 Proses transformasi pemisahan ID

Sebelum dilakukannya perhitungan K-Means dilakukan pengelompokan judul buku, genre, dan penulis menjadi atribut baru yaitu ID. Untuk setiap data judul buku, genre, dan penulis yang unik maka akan menjadi ID yang baru. Pada proses K-means ini menggunakan id untuk menentukan pola peminjaman berdasarkan frekuensi untuk mempermudah dalam perhitungan, Namun sebelum dilakukannya proses visualisasi dan analisis hasil maka perlu dilakukan proses transformasi kembali untuk mengembalikan data id menjadi judul buku, genre, dan penulis. Proses ini dimulai dengan mengambil data keterangan yang ada pada keterangan ID untuk ditambahkan ke data penentuan cluster. Untuk lebih jelasnya dapat dilihat pada kode 4.16 dibawah

```
1. Df_keterangan = pd.read_excel("Keterangan ID.xlsx")
2. df.columns = df.columns.astype(str)
3. df_merged = pd.merge(df, df_keterangan[['id', 'keterangan']],
   on='id')
4. nama_file_excel = "penentuan_cluster.xlsx"
5. df_merged.to_excel(nama_file_excel, index=False)
```

Kode 4. 16 *Pengambilan data keterangan*

Dengan hasil keterangan yang didapatkan melalui kode 4.16 , kemudian data di kolom keterangan di pisahkan atau di split untuk memisahkan antara judul buku, genre, dan penulis menjadi atribut yang baru. Kode 4.17 merupakan kode untuk memisahkan atribut keterangan menjadi judul buku, genre, dan penulis

```
1. df_merged = pd.read_excel("penentuan_cluster.xlsx")
2. df_merged[['Judul Buku', 'Genre Buku', 'Penulis']] =
   df_merged['keterangan'].str.split('_', expand=True)
3. df_merged.drop(columns=['keterangan'], inplace=True)
4. df_merged.columns = df_merged.columns.astype(str)
5. nama_file_excel = "penentuan_cluster.xlsx"
6. df_merged.to_excel(nama_file_excel, index=False)
```

Kode 4. 17 *Split data keterangan*

Dapat dilihat pada kode 4.17 setelah kolom keterangan dipisahkan menjadi judul, genre, dan penulis buku kemudian disimpan kembali di file penentuan cluster dengan format excel. Dari data tersebut kemudian dapat dibaca kelompok clusternya

1. Cluster 1

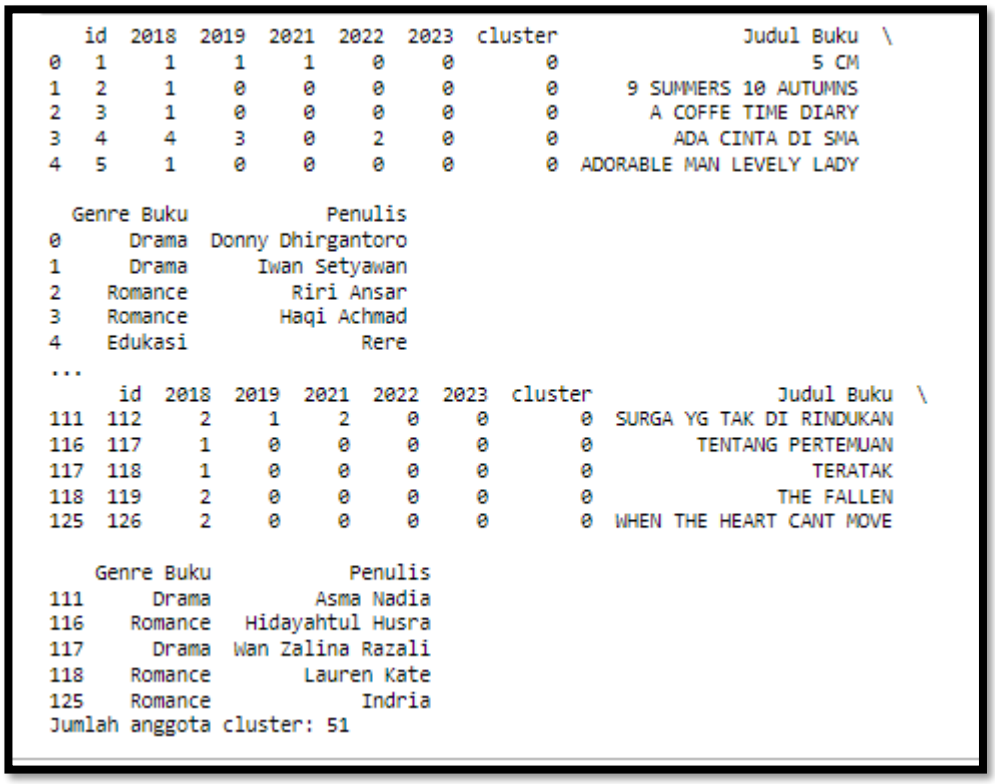
   Karena anggota cluster 1 cukup banyak maka untuk pemanggilan cluster menggunakan kode 4.18 seperti berikut

```
1. df = pd.read_excel("penentuan_cluster.xlsx")
2. first_cluster = df[df['cluster'] == 0]
3. print(first_cluster.head(5))
4. print("...")
5. print(first_cluster.tail(5))
6. print("Jumlah anggota cluster:", first_cluster.shape[0])
```

Kode 4. 18 *Pemanggilan Cluster 1*

Kode di atas dimulai dengan memanggil file excel penentuan cluster , kemudian dari data tersebut dilakukan pemanggilan untuk cluster pertama yang memiliki index 0, karena jumlah cluster cukup banyak maka menggunakan head dan tail yang

digunakan untuk mengambil 5 data diatas dan dibawah untuk ditampilkan. Selain itu juga karena jumlahnya tidak dapat dihitung Secara langsung maka perlu di tampilkan untuk memperjelas jumlah cluster. Adapun untuk anggota cluster 1 dapat dilihat pada gambar 4.10 dibawah



```
    id  2018  2019  2021  2022  2023  cluster                Judul Buku  \
0    1     1     1     1     0     0        0                     5 CM
1    2     1     0     0     0     0        0      9 SUMMERS 10 AUTUMNS
2    3     1     0     0     0     0        0        A COFFE TIME DIARY
3    4     4     3     0     2     0        0          ADA CINTA DI SMA
4    5     1     0     0     0     0        0  ADORABLE MAN LEVELY LADY

  Genre Buku            Penulis
0      Drama  Donny Dhirgantoro
1      Drama      Iwan Setyawan
2    Romance         Riri Ansar
3    Romance        Haqi Achmad
4    Edukasi               Rere
...
      id  2018  2019  2021  2022  2023  cluster                Judul Buku  \
111  112     2     1     2     0     0        0  SURGA YG TAK DI RINDUKAN
116  117     1     0     0     0     0        0         TENTANG PERTEMUAN
117  118     1     0     0     0     0        0                   TERATAK
118  119     2     0     0     0     0        0                THE FALLEN
125  126     2     0     0     0     0        0  WHEN THE HEART CANT MOVE

    Genre Buku            Penulis
111      Drama         Asma Nadia
116    Romance   Hidayahtul Husra
117      Drama  Wan Zalina Razali
118    Romance        Lauren Kate
125    Romance             Indria
Jumlah anggota cluster: 51
```

Gambar 4. 10 *Anggota Cluster 1*

2. Cluster 2

Untuk cluster 2 ini hanya beranggotakan 3 , karena itu menggunakan kode yang berbeda dari 2 cluster lainnya. Adapun untuk kodenya dapat dilihat pada kode 4.19 dibawah.

```
1. df = pd.read_excel("penentuan_cluster.xlsx")
2. second_cluster = df[df['cluster'] == 1]
3. second_cluster_selected = second_cluster[['id', 'Judul Buku',
   'Genre Buku', 'Penulis', '2018', '2019', '2021', '2022',
   '2023']]
4. second_cluster
```

Kode 4. 19 *Pemanggilan Cluster 2*

Untuk kelompok cluster kedua ini dideklarasikan menggunakan second cluster dengan mengambil data cluster 2 yang memiliki index 1 dari penentuan cluster dengan format excel. Untuk anggota cluster 2 ini dapat dilihat pada gambar 4.11 dibawah

	id	2018	2019	2021	2022	2023	cluster	Judul Buku	Genre Buku	Penulis
11	12	14	7	0	0	0	1	AL QURAN XII	Religi	M. Abdul Jalil
12	13	9	6	0	0	0	1	AQIDAH XII	Religi	M. Abdul Jalil
115	116	17	8	3	2	1	1	TARIKH XII	Religi	M. Abdul Jalil

Gambar 4. 11 *Anggota Cluster 2*

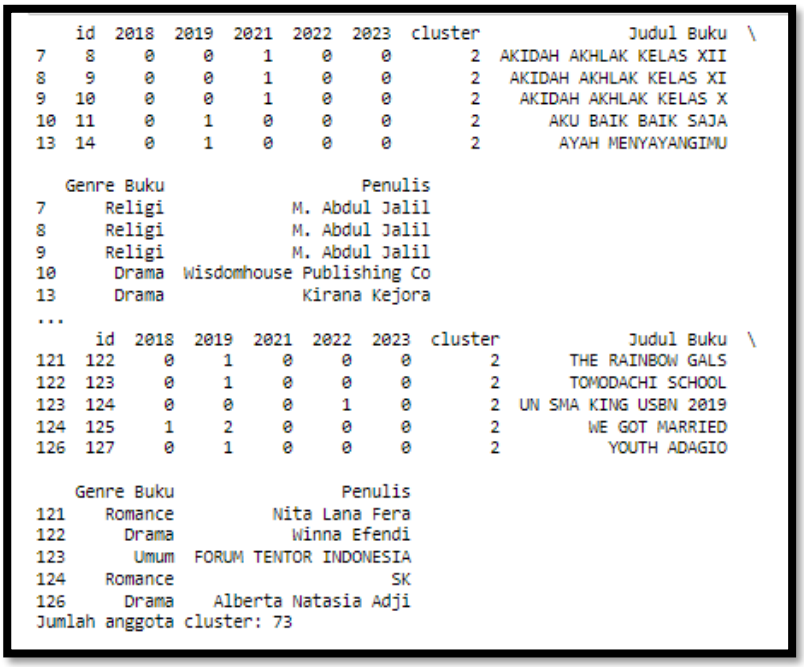
3. Cluster 3

Untuk cluster 3 ini karena memiliki anggota yang cukup banyak seperti cluster 1 maka menggunakan kode yang sama, tetapi dengan index yang berbeda. Untuk lebih jelasnya dapat dilihat pada kode 4.20 dibawah.

```
1. df = pd.read_excel("penentuan_cluster.xlsx")
2. third_cluster = df[df['cluster'] == 2]
3. print(third_cluster.head(5))
4. print("...")
5. print(third_cluster.tail(5))
6. print("Jumlah anggota cluster:", third_cluster.shape[0])
```

Kode 4. 20 *Pemanggilan Cluster 3*

Kode tersebut mirip dengan kode diatasnya. Kode tersebut digunakan untuk memanggil cluster 3 dengan index 2. Adapun untuk anggota clusternya dapat dilihat pada gambar 4.12 dibawah.



```
    id  2018  2019  2021  2022  2023  cluster              Judul Buku  \
7    8     0     0     1     0     0        2  AKIDAH AKHLAK KELAS XII
8    9     0     0     1     0     0        2   AKIDAH AKHLAK KELAS XI
9   10     0     0     1     0     0        2    AKIDAH AKHLAK KELAS X
10  11     0     1     0     0     0        2       AKU BAIK BAIK SAJA
13  14     0     1     0     0     0        2        AYAH MENYAYANGIMU

   Genre Buku                   Penulis
7      Religi             M. Abdul Jalil
8      Religi             M. Abdul Jalil
9      Religi             M. Abdul Jalil
10      Drama  Wisdomhouse Publishing Co
13      Drama              Kirana Kejora
...
      id  2018  2019  2021  2022  2023  cluster            Judul Buku  \
121  122     0     1     0     0     0        2       THE RAINBOW GALS
122  123     0     1     0     0     0        2       TOMODACHI SCHOOL
123  124     0     0     0     1     0        2  UN SMA KING USBN 2019
124  125     1     2     0     0     0        2          WE GOT MARRIED
126  127     0     1     0     0     0        2           YOUTH ADAGIO

    Genre Buku                Penulis
121    Romance          Nita Lana Fera
122      Drama            Winna Efendi
123       Umum  FORUM TENTOR INDONESIA
124    Romance                      SK
126      Drama    Alberta Natasia Adji
Jumlah anggota cluster: 73
```

Gambar 4. 12 *Anggota Cluster 3*

1.3.6 Visualisasi kelompok cluster dan analisis karakteristik tiap cluster

Setelah melakukan klasterisasi data dan pemisahan kolom id menjadi judul , genre dan penulis buku, langkah selanjutnya adalah memvisualisasikan hasil pengelompokan cluster . Kode 4.18 merupakan kode untuk memanggil cluster dan melakukan visualisasi data. Visualisasi yang dihasilkan memberikan pemahaman yang mendalam tentang perilaku peminjaman buku di perpustakaan. Proses dimulai dengan pengelompokan data berdasarkan kluster, di mana nilai maksimum jumlah peminjaman buku per tahun dihitung untuk setiap kluster. Penggunaan plot batang memungkinkan untuk dengan jelas melihat bagaimana pola peminjaman berubah dari tahun ke tahun di setiap kluster. Sumbu x yang mencakup tahun-tahun peminjaman memungkinkan untuk mengikuti perkembangan data peminjaman buku

dari waktu ke waktu, sementara sumbu y yang menunjukkan nilai maksimum peminjaman. Untuk alasan digunakannya nilai maksimum untuk mewakili setiap cluster dalam visualisasi untuk memungkinkan menyoroti titik tertinggi dari aktivitas peminjaman dalam setiap kluster. Dengan demikian, dapat dengan jelas melihat perbedaan dalam intensitas peminjaman antar kluster tanpa risiko adanya tumpang tindih atau tabrakan nilai antar kluster.

Jika menggunakan nilai rata-rata, ada kemungkinan bahwa nilai-nilai akan saling bertabrakan di sepanjang sumbu y, membuat sulit untuk membedakan perbedaan antara kluster yang cukup tipis. Di sisi lain, menggunakan total jumlah peminjaman dapat menyebabkan distorsi jika ada perbedaan signifikan dalam ukuran kluster. Kluster dengan jumlah anggota yang lebih sedikit akan cenderung kalah jumlahnya dari kluster yang lebih besar, meskipun mungkin memiliki tingkat aktivitas peminjaman yang lebih tinggi secara relatif. Oleh karena itu, menggunakan nilai maksimum memberikan solusi yang paling jelas dan representatif untuk menggambarkan aktivitas peminjaman dalam setiap kluster. Kehadiran label dan nilai maksimum di atas setiap batang memperkaya visualisasi dengan memberikan informasi spesifik tentang data. Lebih dari sekadar grafik, visualisasi ini menjadi alat yang kuat untuk mendukung pengambilan keputusan di perpustakaan, dan juga pemetaan yang efisien terhadap koleksi buku. Untuk proses visualisasinya dapat dilihat pada kode 4.21

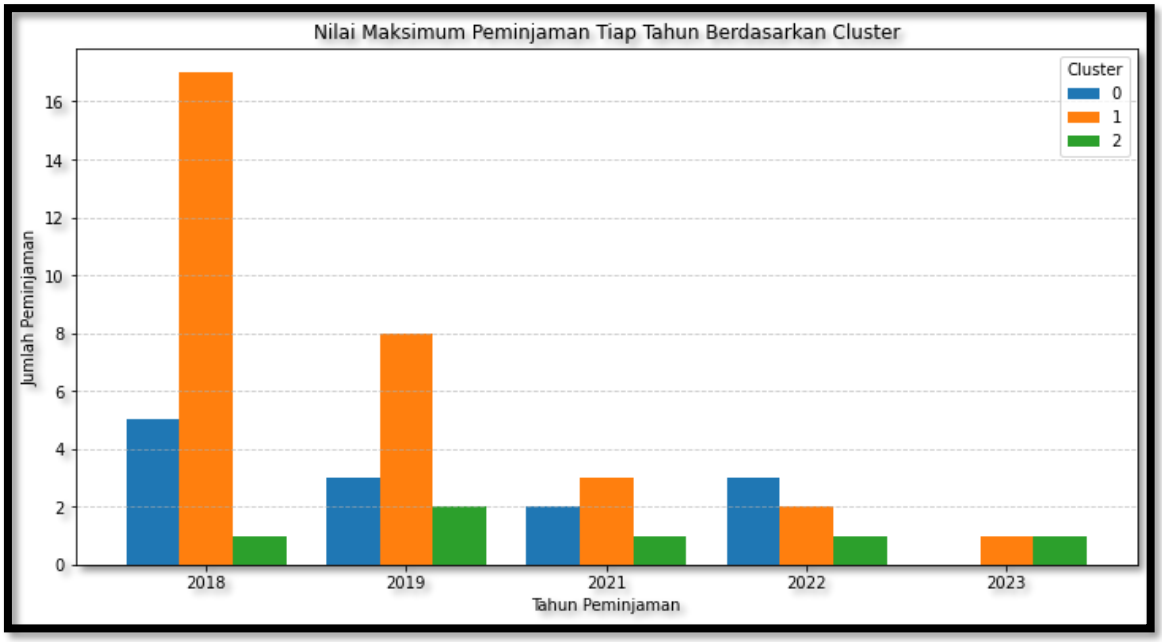
```
df = pd.read_excel("penentuan_cluster.xlsx")
tahun_cluster = df[['2018', '2019', '2021', '2022', '2023',
'cluster']]
max_yearly_counts = tahun_cluster.groupby('cluster').max()
max_yearly_counts.T.plot(kind='bar', figsize=(12, 6), width=0.8)
plt.title('Nilai Maksimum Peminjaman Tiap Tahun Berdasarkan
Cluster')
plt.xlabel('Tahun Peminjaman')
plt.ylabel('Jumlah Peminjaman')
plt.xticks(rotation=0)
plt.legend(title='Cluster')
      plt.grid(axis='y', linestyle='--', alpha=0.7)
      plt.show()
```

Kode 4. 21 *Visualisasi Cluster*

Kode ini digunakan untuk membuat visualisasi yang menampilkan nilai tertinggi dari setiap tahun peminjaman buku dalam kluster-kluster yang dihasilkan dari proses klasterisasi. Pertama, data dibaca dari file Excel "penentuan_cluster.xlsx" dan disimpan dalam DataFrame df. Kolom-kolom yang relevan ('2018', '2019', '2021', '2022', '2023', dan 'cluster') dipilih dan disimpan dalam DataFrame tahun_cluster. Data ini kemudian dikelompokkan berdasarkan kluster menggunakan metode groupby, dan nilai maksimum dari setiap kluster dihitung untuk setiap tahun. Hasil pengelompokan ini disimpan dalam DataFrame max_yearly_counts.

Selanjutnya, data ini diputar (transpose) sehingga kolom tahun menjadi sumbu x dan kluster menjadi legenda. Plot batang dibuat menggunakan plt.bar, dengan ukuran 12x6 inci dan lebar batang 0.8. Judul grafik diatur menjadi "Nilai Maksimum Peminjaman Tiap Tahun Berdasarkan Cluster". Label sumbu x diatur untuk menunjukkan tahun-tahun peminjaman, sementara sumbu y menunjukkan jumlah peminjaman tertinggi per tahun. Label sumbu x diputar 0 derajat untuk tampilan yang lebih rapi. Legenda diberi judul "Cluster". Garis grid horizontal ditambahkan dengan gaya garis putus-putus dan transparansi 0.7 untuk memudahkan pembacaan nilai pada grafik. Terakhir, grafik tersebut ditampilkan menggunakan

plt.show() . Visualisasi ini membantu dalam memahami tren maksimum peminjaman buku per tahun di setiap cluster. Gambar 4.13 merupakan hasil visualisasi dari kode diatas.



Gambar 4. 13 *Visualisasi Pola Cluster*

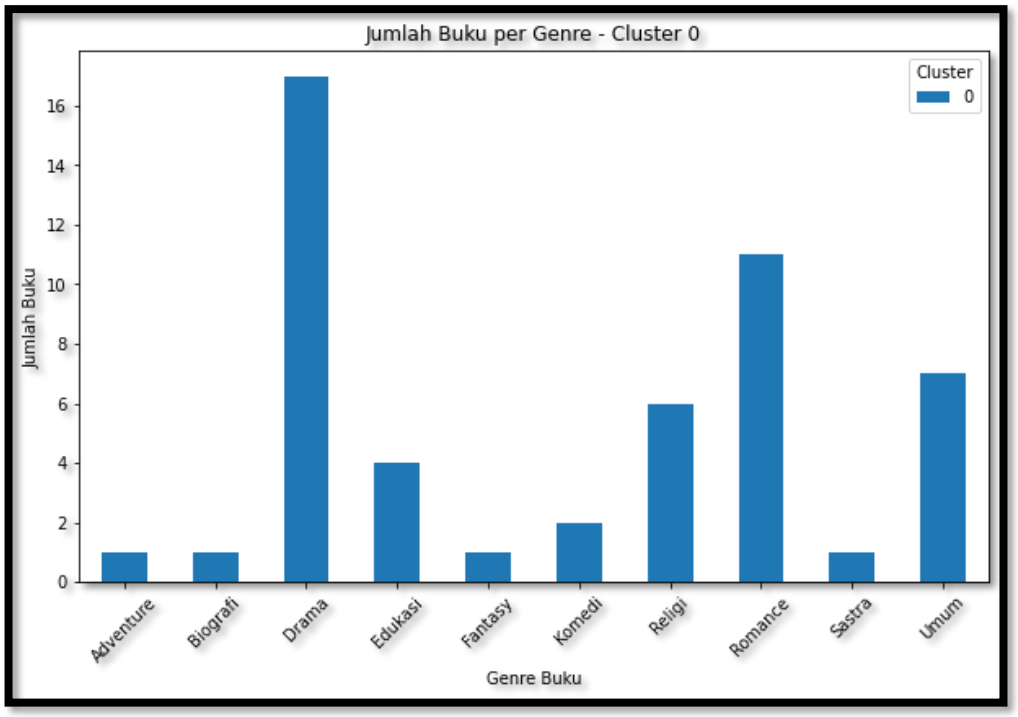
Berdasarkan gambar 4.13 visualisasi pola cluster ini juga dapat digunakan untuk melihat pola genre buku yang dibaca berdasarkan kelompok clusternya. Dengan melihat ini dapat diketahui juga karakteristik tiap cluster berdasarkan genre buku. Genre buku ini dapat digunakan untuk mengetahui karakteristik karena tidak terlalu bervariasi dan memungkinkan untuk divisualisasi. Kode 4.22 Merupakan kode untuk visualiasi genre buku tiap cluster

```
df = pd.read_excel("penentuan_cluster.xlsx")
genre_counts = df.groupby(['cluster', 'Genre Buku']).size().unstack(fill_value=0)
for cluster in genre_counts.index:
    genre_counts.loc[cluster].plot(kind='bar', stacked=True, figsize=(10,6))
    plt.title(f'Jumlah Buku per Genre - Cluster {cluster}')
    plt.xlabel('Genre Buku')
    plt.ylabel('Jumlah Buku')
    plt.xticks(rotation=45)
    plt.legend(title='Cluster')
        plt.show()
```

Kode 4. 22 *Persebaran data berdasar genre buku*

1. Cluster 1

   Untuk hasil visualisasi cluster 1 berdasarkan genre dalam dilihat pada gambar 4.14 dibawah.

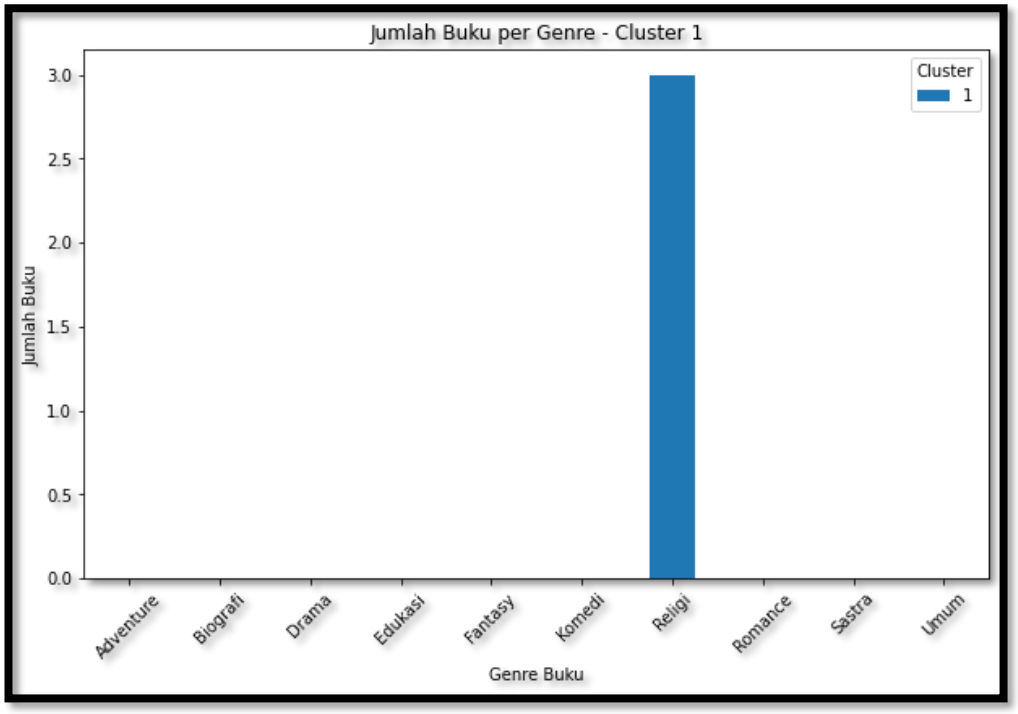


Gambar 4. 14 *Persebaran Genre Cluster 1*

Dari visualisasi tersebut dapat diketahui bahwa cluster 1 ini didominasi oleh buku dengan genre drama, dan romance. Meskipun demikian genre buku dalam cluster ini cukup bervariasi. Membuat data terlihat tidak terlalu didominasi oleh genre tertentu karena selisih antar jumlah genre yang tidak terlalu besar.

2. Cluster 2

Untuk hasil visualisasi cluster 2 berdasarkan genre dalam dilihat pada gambar 4.15 dibawah.

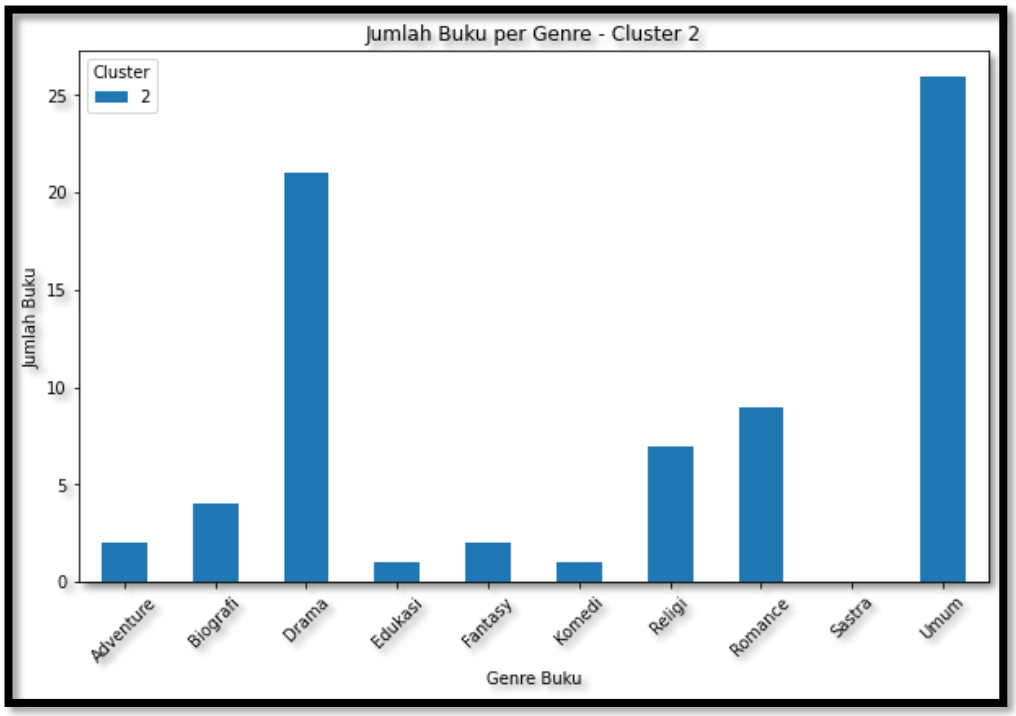


Gambar 4. 15 *Persebaran Genre Cluster 2*

Untuk cluster kedua ini buku yang dipinjam hanya yang bergenre religi. Hal ini menjadikan genre buku pada cluster ini kurang bervariasi jika dibandingkan dengan cluster lainnya karena sangat didominasi pada genre religi. Tentu hal ini mempermudah untuk proses presentasi pengetahuan akan tetapi kurang menarik jika melihat dari hasil visualisasi datanya

3. Cluster 3

Untuk hasil visualisasi cluster 3 berdasarkan genre dalam dilihat pada gambar 4.16 dibawah.



Gambar 4. 16 *Persebaran Genre Cluster 3*

Dari visualisasi tersebut dapat diketahui bahwa cluster 3 ini didominasi oleh buku dengan genre Umum dan Drama. Genre genre buku pada cluster ini cukup bervariasi seperti cluster 1 . Meskipun demikian selisih jumlah genre buku yang dipinjam cukup jauh satu sama lain dengan memperlihatkan bahwa buku angka peminjaman didominasi oleh 2 genre. Jumlah dari 2 genre ini jauh dari jumlah peminjaman genre lain, meskipun genre romance adalah genre ketiga terbanyak akan tetapi memiliki jumlah yang jauh jika dibandingkan dengan genre drama yang berada diposisi 2 yang memiliki selisih lebih dari 10.

1.3.7 Representasi pengetahuan dan pembahasan

Dari Gambar 4.12, Penelitian menggunakan metode K-means clustering ini membagi buku-buku ke dalam tiga cluster yang masing-masing memiliki karakteristik unik. Evaluasi pengetahuan dari hasil clustering ini akan membantu pembaca memahami perbedaan antara kluster-kluster tersebut serta memberikan wawasan mengenai preferensi pembaca perpustakaan.

1. Cluster 1

Kluster 1 terdiri dari buku-buku yang cukup diminati, meskipun popularitasnya tidak sebesar kluster 2. Buku-buku dalam kluster ini memiliki daya tarik yang konsisten dan berpotensi untuk berkembang di masa mendatang. Tren peminjaman buku dalam kluster ini stabil atau sedikit fluktuatif dari waktu ke waktu, menunjukkan minat yang cukup konstan dari pembaca. Genre yang dominan dalam kluster ini adalah drama dan romance, yang menunjukkan bahwa pembaca masih tertarik pada cerita fiksi dengan plot yang menarik dan karakter yang kuat. Meskipun buku-buku dalam kluster ini tidak mencapai popularitas puncak, mereka memiliki basis pembaca yang setia dan berpotensi untuk menarik lebih banyak minat dengan promosi yang tepat atau penambahan koleksi baru dalam genre yang sama.

2. Cluster 2

Cluster 2 merupakan kategori buku yang paling diminati di perpustakaan, dengan jumlah peminjaman tertinggi dibandingkan dengan kluster lainnya. Buku-buku dalam kluster ini sangat diminati dan menjadi favorit utama pengunjung perpustakaan. Daya tarik yang kuat dan perhatian signifikan dari pembaca membuat kluster ini menonjol.

Buku-buku dalam kluster ini sebagian besar bergenre religi dan ditulis oleh penulis terkenal seperti M Abdul Jalil, menunjukkan bahwa tema-tema spiritual dan karya dari penulis populer memiliki daya tarik besar. Popularitas buku-buku ini bisa menjadi indikasi bagi perpustakaan untuk memperbanyak koleksi dalam genre ini dan menambahkan karya dari penulis ini untuk terus menarik minat pembaca.

3. Cluster 3:

Kluster 3 menunjukkan kategori buku yang kurang diminati, meskipun masih dipinjam setiap tahun. Jumlah peminjaman dalam kluster ini cenderung stagnan dan tidak mengalami peningkatan signifikan dari tahun ke tahun. Buku-buku dalam kluster ini mungkin memiliki minat yang lebih terbatas atau khusus, sehingga tidak mendapat perhatian besar dari pengunjung perpustakaan. Genre yang dominan dalam kluster ini adalah drama dan umum yang kurang populer, menunjukkan bahwa cerita-cerita dalam genre ini mungkin tidak sejalan dengan minat utama pembaca saat ini. Untuk meningkatkan popularitas buku-buku dalam kluster ini, perpustakaan mungkin perlu mengevaluasi ulang koleksi dalam genre ini dan mempertimbangkan untuk menambahkan buku-buku yang lebih relevan atau menarik bagi pembaca.

Dengan hasil evaluasi dari system yang telah dibuat ini,diharapkan dapat membantu pihak perpustakaan membuat keputusan yang lebih baik dalam mengelola koleksi buku. Tujuannya adalah untuk meningkatkan minat pembaca melalui pengadaan buku, memastikan bahwa perpustakaan tetap relevan dan menarik bagi pengunjungnya. Informasi dari kluster ini juga bisa digunakan untuk mengidentifikasi area di mana koleksi dapat ditingkatkan atau diperluas sesuai dengan minat dan kebutuhan pembaca.

BAB V. Kesimpulan Dan Saran

2.1 Kesimpulan

Penelitian ini bertujuan untuk mengidentifikasi pola peminjaman buku di perpustakaan SMA 5 Muhammadiyah setelah terjadi penurunan akibat pandemi COVID-19, dengan fokus pada penggunaan metode K-Means Clustering. Dalam mencapai tujuan tersebut, langkah-langkah yang dilakukan meliputi pengumpulan dan preprocessing data peminjaman buku, normalisasi data untuk memastikan konsistensi skala fitur, serta pemilihan jumlah klaster optimal berdasarkan evaluasi menggunakan Silhouette Score, Davies-Bouldin Score, dan Calinski-Harabasz Score.

Hasil pengelompokan menunjukkan bahwa terdapat tiga pola utama peminjaman buku di perpustakaan, yaitu buku yang paling diminati (genre religi), cukup diminati (genre drama dan romance), dan kurang diminati (genre drama dan Umum). Evaluasi kualitas cluster dengan matriks evaluasi tersebut menunjukkan bahwa pengelompokan memiliki kualitas yang baik dengan nilai Silhouette Score mendekati 0,791, Davies-Bouldin Score sebesar 0,626, dan Calinski-Harabasz Score sebesar 615,093. Dari hasil analisis pola peminjaman buku ini, disarankan agar pihak perpustakaan SMA 5 Muhammadiyah dapat mengambil langkah-langkah strategis berupa:

1. **Promosi Intensif Buku Kurang Diminati:** Melakukan promosi khusus dan acara untuk meningkatkan minat baca pada buku-buku dalam klaster yang kurang diminati, seperti dengan memperluas dan memperbarui koleksi yang relevan.
2. **Diversifikasi Koleksi Buku:** Memperbanyak koleksi buku dalam klaster yang cukup diminati untuk mempertahankan minat pembaca, terutama pada genre drama dan romance yang menunjukkan tren peminjaman yang stabil.

3. **Pemantauan dan Evaluasi Berkala:** Melakukan evaluasi secara berkala terhadap pola peminjaman buku untuk memastikan strategi yang diambil efektif dan dapat disesuaikan dengan perubahan minat pembaca.

Dengan implementasi langkah-langkah ini, diharapkan pihak perpustakaan dapat meningkatkan minat baca siswa dan efektivitas pengelolaan koleksi buku secara signifikan.

2.2 Saran

Guna memperoleh hasil yang semakin menguatkan penelitian ini, maka terdapat sejumlah saran untuk pengembangan lebih lanjut bagi penelitian slanjutnya seperti berikut:

1. Menggunakan data yang Lebih Luas dengan mengumpulkan data peminjaman buku dari berbagai perpustakaan dengan ukuran dan karakteristik yang berbeda untuk meningkatkan generalisasi hasil penelitian. Variasi data ini dapat membantu dalam memahami pola peminjaman yang berbeda di berbagai jenis perpustakaan.
2. Mencoba metode clustering lain seperti Hierarchical Clustering, DBSCAN, atau Mean Shift dengan berdasarkan kasus yang serupa tetapi dengan kebutuhan yang berbeda.
3. Menambahkan lebih banyak fitur ke dalam analisis, atribut lain untuk melakukan proses kmeans cluster untuk memperoleh pemahaman yang lebih mendalam tentang perilaku peminjaman. Dengan demikian membuat penelitian memiliki segmentasi yang lebih spesifik dan berguna.
4. Selain metrik evaluasi yang digunakan, mencoba metrik lain seperti Adjusted Rand Index (ARI) atau Mutual Information Score untuk mengevaluasi hasil clustering. Ini dapat memberikan wawasan tambahan mengenai kualitas klaster yang dihasilkan.
5. Menggunakan visualisasi grafik lain, seperti scatter plot atau heatmap, untuk memperlihatkan perbandingan antara metrik evaluasi yang digunakan

Daftar Pustaka

[1] I. A. Nur Afifah and H. Nurdiyanto, "Data Mining Clustering Dalam Pengelompokan Buku Perpustakaan Mengunakan Algoritma K-Means," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 8, no. 3, pp. 802–814, 2023, doi: 10.29100/jipi.v8i3.3891.

[2] W. A. F. Dewi, "Dampak COVID-19 terhadap Implementasi Pembelajaran Daring di Sekolah Dasar. Edukatif : Jurnal Ilmu Pendidikan, 2(1), 55–61. https://doi.org/10.31," *Edukatif J. Ilmu Pendidik.*, vol. 2, no. 1, pp. 55–61, 2020.

[3] I. M. Haryani, Dicky Nofriansyah, "Implementasi Data Mining Untuk Pengelompokan Buku Di Perpustakaan Yayasan Nurul Islam Indonesia Baru Dengan Metode K-Means Clustering," *J. Cyber TechTech*, vol. 1, no. 1, pp. 1–12, 2021.

[4] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*. 2011.

[5] J. Nasir, "Penerapan Data Mining Clustering Dalam Mengelompokan Buku Dengan Metode K-Means," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 11, no. 2, pp. 690–703, 2021, doi: 10.24176/simet.v11i2.5482.

[6] E. Bakker, "Implementasi Data Mining Clustering Data Perpustakaan Menggunakan Algoritma K-Means untuk Menentukan Penambahan Koleksi Buku di Perpustakaan UPY," *Semin. Nas. Din. Inform.*, vol. 7, no. 1, pp. 22–25, 2020.

[7] S. Butsianto and N. T. Mayangwulan, "Penerapan Data Mining Untuk Prediksi Penjualan Mobil Menggunakan Metode K-Means Clustering," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 3, no. 3, pp. 187–201, 2020, doi: 10.32672/jnkti.v3i3.2428.

[8] A. Nugraha, O. Nurdiawan, and G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Yana Sport," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 849–855, 2022, doi: 10.36040/jati.v6i2.5755.

[9] A. Febriyanto, S. Achmadi, and A. P. Sasmito, "Penerapan Metode K-Means Untuk Clustering Pengunjung Perpustakaan Itn Malang," *J. Mhs. Tek. Inform.*, vol. 5, no. 1, pp. 61–70, 2021.

[10] N. A. Rahmalinda and A. Jananto, "Penerapan Metode K-Means Clustering Dalam Menentukan Strategi Promosi Berdasarkan Data Penerimaan Mahasiswa Baru," *J. Tekno Kompak*, vol. 16, no. 2, pp. 163–175, 2022.

[11] R. Amalia, "Penerapan Data Mining untuk Memprediksi Hasil Kelulusan Siswa Menggunakan Metode Naïve Bayes," *Juisi*, vol. 06, no. 01, pp. 33–42, 2020.

[12] jagoanhosting.com, "Data Mining: Pengertian, Fungsi, Metode & Penerapannya," jagoanhosting.com. Accessed: Jun. 02, 2024. [Online]. Available: https://www.jagoanhosting.com/blog/apa-itu-data-mining/

[13] sis.binus.ac.id, "Proses Data Mining KDD," sis.binus.ac.id. Accessed: Jun. 02, 2024. [Online]. Available: https://sis.binus.ac.id/2021/09/30/proses-data-mining-kdd/

[14] G. E. I. Kambey, R. Sengkey, and A. Jacobus, "Penerapan Clustering pada Aplikasi Pendeteksi Kemiripan Dokumen Teks Bahasa Indonesia," *J. Tek. Inform.*, vol. 15, no. 2, pp. 75–82, 2020, [Online]. Available: https://ejournal.unsrat.ac.id/v3/index.php/informatika/article/view/28907/28665

[15] Spada.uns.ac.id, "Algoritma K-Means," *spada.uns.ac.id*. 2023. [Online]. Available: https://spada.uns.ac.id/pluginfile.php/871813/mod_resource/content/1/Materi clustering.pdf

[16] sis.binus.ac.id, "Clustering Algoritma (K-Means)," sis.binus.ac.id. Accessed: Jun. 02, 2024. [Online]. Available: https://sis.binus.ac.id/2022/01/31/clustering-algoritma-k-means/

[17] klc2.kemenkeu.go.id, "Koefisien Silhouette untuk Menentukan Jumlah Klaster Ideal," klc2.kemenkeu.go.id. Accessed: Jun. 02, 2024. [Online]. Available: https://klc2.kemenkeu.go.id/kms/knowledge/koefisien-silhouette-untuk-menentukan-jumlah-klaster-ideal-d0e05a09/detail/

[18] techtarget.com, "Data Visualization," techtarget.com. Accessed: Jun. 06, 2024. [Online].

Available: https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization
[19] Y. Galahartlambang, T. Khotiah, and J. Jumain, "Visualisasi Data Dari Dataset COVID-19 Menggunakan Pemrograman Python," *J. Ilm. Intech Inf. Technol. J. UMUS*, vol. 3, no. 01, pp. 58–64, 2021, [Online]. Available: http://jurnal.umus.ac.id/index.php/intech/article/view/417
[20] S. Ayu, Anggi; Nugroho, Nur; Muhammad, "Penerapan Data Mining Untuk Menentukan Persediaan Barang Pada PT. Deli Food Menggunakan Metode K-Means," vol. 4, no. 7, 2021, [Online]. Available: https://ojs.trigunadharma.ac.id/
[21] J. S. Sibatuara, "Klasterisasi Tingkat Pemahaman Siswa Dalam Sistem Pembelajaran Online Dengan Metode K-Means Clustering Skripsi Fakultas Teknik Universitas Medan Area Medan Klasterisasi Tingkat Pemahaman Siswa Dalam Sistem Pembelajaran Online Dengan Metode K-Means Cluster," 2021.
[22] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA Tek. J. Ilm. Ilmu-Ilmu Tek.*, vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.
[23] Teradata.com, "What Is Data Cleansing?," teradata.com. Accessed: Jun. 02, 2024. [Online]. Available: https://www.teradata.com/insights/data-platform/what-is-data-cleansing
[24] S. Pujiono, R. Astuti, and F. Muhamad Basysyar, "Implementasi Data Mining Untuk Menentukan Pola Penjualan Produk Menggunakan Algoritma K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 615–620, 2024, doi: 10.36040/jati.v8i1.8360.
[25] mie.binus.ac.id, "Teknik pre-processing dan classification dalam data science," mie.binus.ac.id. Accessed: Jun. 02, 2024. [Online]. Available: https://mie.binus.ac.id/2022/08/26/teknik-pre-processing-dan-classification-dalam-data-science/
[26] algorit.ma, "Pengenalan Library Python," algorit.ma. Accessed: Jun. 02, 2024. [Online]. Available: https://algorit.ma/blog/library-python/
[27] pandas.pydata.org, "How Pandas Read Excel," pandas.pydata.org. Accessed: Jun. 02, 2024. [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.read_excel.html
[28] Atlan.com, "Standardize Data: Why It Matters & How to Do It Effectively!," Atlan.com. Accessed: Jun. 02, 2024. [Online]. Available: https://atlan.com/standardize-data/