

CLUSTERING MENGGUNAKAN METODE K-MEANS UNTUK MENENTUKAN SISWA BERPRESTASI



**UNIVERSITAS
TEKNOLOGI
SUMBAWA**

Diampu oleh : Herfandi, A.Md, S.Kom, M.Kom

Disusun oleh : Kelompok 3 Jam 1

Ketua : Abigail Perkasa

Anggota : Cece Lestiana

Dimas Arya Mukhti

Sahara Hasibuan

Tyreenia

Kelas : Artificial Intelligence (B)

UNIVERSITAS TEKNOLOGI SUMBAWA

2021

**Alamat : Jl. Raya Olat Maras, BatuAlang, Moyo Hulu, Pernek, Moyohulu,
Kabupaten Sumbawa, Nusa Tenggara Barat. 84371, Indonesia
Telp. (0371) 2629009 Email : informasi@uts.ac.id**

KATA PENGANTAR

Segala puji hanya milik Allah SWT, shalawat serta salam selalu tercurahkan kepada Rasulullah SAW. Berkat limpahan rahmat-Nya, kami mampu menyelesaikan tugas makalah ini guna memenuhi tugas mata kuliah Artificial Intelligence. Dalam penyusunan tugas atau materi ini, tidak sedikit hambatan yang penulis hadapi, namun penulis menyadari bahwa kelancaran dalam penyusunan materi ini tidak lain berkat bantuan, dorongan dan bimbingan orang tua serta bapak Herfandi selaku dosen mata kuliah Artificial Intelligence sehingga kendala-kendala yang penulis hadapi dapat teratasi.

Karya ilmiah disusun agar pembaca dapat memperluas wawasan mengenai penelitian menggunakan metode pendekatan K-Means yang kami sajikan berdasarkan pengamatan dari berbagai sumber informasi, referensi dan contoh soal. Makalah ini penulis susun dengan berbagai rintangan baik itu yang datang dari diri kami maupun yang datang dari luar. Namun dengan penuh kesabaran dan terutama pertolongan dari Allah akhirnya karya ilmiah ini dapat terselesaikan.

Semoga makalah ini dapat memberikan wawasan yang lebih luas dan menjadi sumbangan pemikiran kepada pembaca khususnya para mahasiswa Universitas Teknologi Sumbawa. Kami sadar bahwa karya ilmiah ini masih banyak kekurangan dan jauh dari sempurna. Untuk itu, kepada dosen pengampu, penulis meminta masukannya demi perbaikan pembuatan karya ilmiah kami di masa yang akan datang dan mengharapkan kritik dan saran dari para pembaca.

Sumbawa Besar, 20 Desember 2021

Penyusun

ABSTRACT

Student data continues to grow every year and produces abundant data so that data accumulation occurs. Abundant data need to do data processing to explore the information contained in the data. The purpose of this study is to cluster student data through a data mining process using the K-Means algorithm for cluster formation. Attribute data used is the name of the student, NIS, the value of the subject. The data used is student data with a sample of 107 items and the data source comes from the website. There were 5 student clusters, namely Cluster 1 (cluster_0) 19 items, Cluster 2 (cluster_1) 24 items, Cluster 3 (cluster_2) 20 items, Cluster 4 (cluster_3) 26 items, and Cluster 5 (cluster_4) 17 items. The results of this study are used as a basis for determining students who excel in this data based on the results of clustering.

Keywords: K-Means, Clustering, Data Mining, Data, Cluster

ABSTRAK

Data siswa setiap tahunnya terus bertambah dan menghasilkan data yang berlimpah sehingga terjadi penumpukan data. Data yang berlimpah perlu dilakukan pengolahan data untuk menggali informasi yang terdapat didalam data tersebut. Tujuan penelitian ini untuk mengkluster data siswa melalui proses data mining dengan menggunakan algoritma K-Means untuk pembentukan cluster. Atribut data digunakan adalah nama siswa, NIS, nilai mata pelajaran. Data yang digunakan adalah data siswa dengan sampel data 107 items dan sumber data berasal dari website. Cluster siswa yang terbentuk ada 5 yaitu Cluster 1 (cluster_0) 19 items, Cluster 2 (cluster_1) 24 items, Cluster 3 (cluster_2) 20 items, Cluster 4 (cluster_3) 26 items, dan Cluster 5 (cluster_4) 17 items. Hasil dari penelitian ini di gunakan sebagai dasar untuk menentukan siswa yang berprestasi yang ada di data ini berdasarkan hasil clustering.

Keywords: K-Means, Clustering, Data Mining, Data, Cluster

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Tingginya prestasi murid dan minimnya angka murid yang tidak berprestasi mencerminkan keunggulan sektor pendidikan. Sektor Pendidikan kini diharapkan mampu bersaing dengan memanfaatkan kemajuan SI/TI, yang dapat menunjang peningkatan daya saing dan menunjang operasional sehari-hari serta pengambilan keputusan strategis. Secara umum, keberhasilan murid dievaluasi berdasarkan evaluasi pelajaran teoritis dan praktis, serta kehadiran dan ketidakhadiran murid selama di dalam kelas.

Penilaian dibagi menjadi tiga kategori yaitu pengetahuan, bakat, dan sikap. Pengajar menilai semua murid yang mengikuti pelajaran yang diberikan guna mengevaluasi dan menganalisis prestasi belajar murid. Jumlah data murid terus bertambah setiap tahunnya. Hal ini mengakibatkan akumulasi data yang belum diolah secara baik digunakan untuk mengungkap pengetahuan dan informasi baru dengan pattern yang terbentuk sebagai hasil akumulasi data. Jumlah pemrosesan data yang terus meningkat mengharuskan penggunaan strategi dan metode sehingga dapat ditransformasikan menjadi informasi dan pengetahuan yang dapat dimanfaatkan oleh pendidik dalam proses pembuatan kebijakan. Hal ini menjadi permasalahan bagi guru maupun wali kelas yaitu bagaimana cara menentukan tingkat prestasi murid yang rendah, cukup, dan tinggi serta menemukan top rank murid unggulan dalam kelas, agar dapat membentuk kelas yang ideal untuk meningkatkan prestasi maupun memotivasi murid.

Data mining merupakan teknik untuk mengungkap tautan, pattern, dan tren baru dengan memfilter kumpulan banyak data yang disimpan di penyimpanan dan menerapkan teknik pengenalan pattern seperti prosedur statistik dan matematik. K-Means adalah algoritma data mining yang dapat digunakan untuk mengelompokkan/mengklasterkan data yang sangat besar atau bertumpuk yang dapat ditangani dengan salah satu dari beberapa cara, termasuk clustering. Beberapa penelitian tentang penerapan pendekatan Means Clustering telah dilakukan untuk menganalisis prestasi belajar murid yang menghasilkan kelompok murid berprestasi. Pengelompokan tersebut didasarkan data nilai tes, Tes Tengah Semester (TTS), Tes

Akhir Semester (TAS), maupun keaktifan absensi murid di beberapa sekolah. Pendekatan K-Means digunakan untuk rekomendasi jurusan calon mahasiswa berdasarkan nilai untuk pengelompokan data pemilihan jurusan bagi Perguruan Tinggi. Fokus pada penelitian ini adalah pada pengelompokan murid berdasarkan nilai mata pelajaran dan absensi menjadi kategori tinggi, cukup, dan rendah menggunakan metode K-Means Clustering. Selanjutnya dari cluster tinggi akan dicari siswa unggul menggunakan metode Simple Additive Weighting (SAW). Dalam penelitian ini akan dibahas empat bagian yaitu bagian 1 berisi pendahuluan yang menjelaskan mengenai objek penelitian, permasalahan, metode yang digunakan dan manfaat yang mengangkat topik clustering K-Means dan teori tentang metode K-Means. Pada bagian 2 berisi landasan teori yang digunakan dalam melakukan penelitian. Selanjutnya, bagian 3 berisi metodologi penelitian mengenai topik clustering dengan metode k-means. Pada bagian 4 berisi tentang pembahasan hasil clustering prestasi para murid, menggunakan algoritma K-Means. Pada bagian 5 berisi kesimpulan dan saran serta daftar pustaka.

1.2 Rumusan Masalah

Bagaimana menggunakan pendekatan clustering k-means untuk mendapatkan data siswa yang berprestasi?

1.3 Batasan Masalah

Pembatasan suatu masalah digunakan untuk menghindari adanya penyimpangan maupun pelebaran pokok masalah agar penelitian tersebut lebih terarah dan memudahkan dalam pembahasan sehingga tujuan penelitian akan tercapai. Beberapa batasan masalah dalam penelitian ini adalah sebagai berikut:

- 1.3.1 Luas lingkup hanya meliputi informasi seputar algoritma K-Means.
- 1.3.2 Informasi yang disajikan yaitu: pengertian Data Mining, Clustering, K-Means beserta bagaimana pengimplementasiannya untuk menemukan data siswa yang berprestasi.

1.4 Tujuan

1.4.1 Tujuan Umum

Untuk mengetahui segala sesuatu yang berkaitan dengan clustering k-means.

1.4.2 Tujuan Khusus

Untuk mengetahui bagaimana cara menggunakan pendekatan clustering k-means untuk mendapatkan data siswa yang berprestasi.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dengan tercapainya tujuan penelitian ini diantaranya yaitu :

1.5.1 Manfaat Bagi Penulis

Sebagai sarana untuk menerapkan ilmu yang telah didapatkan selama perkuliahan sekaligus untuk memenuhi UAS mata kuliah *Artificial Intelligence*.

1.5.2 Manfaat Bagi Pembaca

Dapat digunakan sebagai informasi dan tambahan pengetahuan tentang algoritma klustering.

1.6 Sistematika Penulisan

Sistematika penulisan tugas akhir ini disusun untuk memberikan gambaran umum tentang penelitian yang dijalankan. Sistematika penulisan tugas akhir ini adalah sebagai berikut:

BAB I. PENDAHULUAN

Bab ini berisi latar belakang masalah, identifikasi masalah, maksud dan tujuan yang ingin dicapai, batasan masalah, metodologi penelitian yang diterapkan dalam memperoleh dan mengumpulkan data serta sistematika penulisan.

BAB II. LANDASAN TEORI

Menjelaskan tentang kajian pustaka serta teori yang melandasi penelitian algoritma clustering untuk menentukan siswa berprestasi.

BAB III. METODE PENELITIAN

Menjelaskan tentang metode penelitian dari pengumpulan data eksperimen dengan menguji data yang ada menggunakan algoritma clustering yang memprediksi siswa berprestasi.

BAB IV. HASIL DAN PEMBAHASAN

Menjelaskan dan menampilkan hasil prediksi dengan menggunakan algoritma clustering.

BAB V. PENUTUP

Berisi kesimpulan dari implementasi dan uji coba yang dilakukan. Selain itu berisi pula saran yang diharapkan dapat menjadi masukan untuk pengembangan di masa datang.

BAB 2

LANDASAN TEORI

2.1 Penelitian Terdahulu

No	Pengarang	Judul Jurnal	Pembahasan
1	Green F Mandias, Green A Sandag, Susi, Haryanto (2017)	Penerapan Algoritma K-Means Untuk Analisis Prestasi Akademik Mahasiswa Fakultas Ilmu Komputer Universitas Klabat	Menghasilkan output 3 cluster yang berisi tentang berapa persen mahasiswa dengan nilai tinggi, sedang, dan rendah

2	Koko Handoko (2016)	<p>Penerapan data mining dalam Meningkatkan mutu pembelajaran pada</p> <p>Instansi perguruan tinggi menggunakan Metode k-means clustering (studi kasus di Program studi tkj akademi komunitas Solok selatan)</p>	Dapat menghasilkan data mutu pembelajaran yang dilihat dari cluster
3	Yohanni Sahra (2018)	<p>Penerapan Data Mining Dalam Pengelompokkan Data</p> <p>Nilai Siswa Untuk</p> <p>Penentuan Jurusan</p> <p>Siswa Pada SMA</p> <p>Tamora Menggunakan</p> <p>Algoritma K-Means Clustering</p>	dapat digunakan dalam pengelompokkan data nilai siswa untuk penentuan jurusan pada SMA Tamora

4	Gede Aditra Pradyana, Agus Aan Jiwa Permana (2018)	Sistem pembagian kelas kuliah mahasiswa dengan metode k-means dan knearest neighbors	Dari metode tersebut kemudian dikembangkan sebuah sistem berbasis web dengan menggunakan Bahasa PHP dengan framework Laravel
---	-------------------------------------------------------------------	--------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------

2.2 Landasan Teori

2.2.1 Data Mining

Perkembangan teknologi informasi telah memberikan kontribusi pada cepatnya pertumbuhan jumlah data yang dikumpulkan dan disimpan dalam basis data berukuran besar (big data). Big data adalah istilah yang menggambarkan volume data yang besar, baik data yang terstruktur maupun data yang tidak terstruktur. Big data memiliki potensi tinggi untuk mengumpulkan wawasan kunci dari informasi bisnis. Big data dapat dianalisis untuk wawasan yang mengarah pada pengambilan keputusan dan strategi bisnis yang lebih baik.

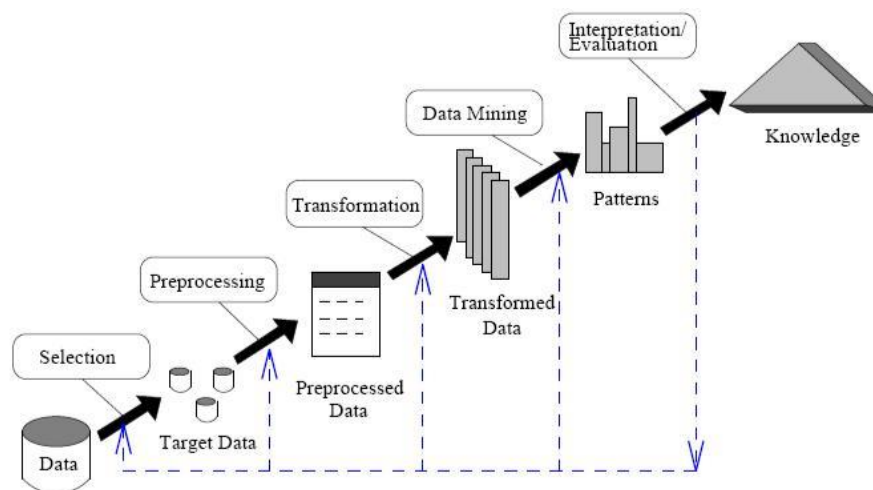
Sebuah metode atau teknik diperlukan untuk dapat merubah data tersebut menjadi sebuah informasi berharga atau pengetahuan yang bermanfaat untuk mendukung pengambilan keputusan. Suatu teknologi yang dapat digunakan untuk mewujudkannya adalah data mining. Belakangan ini data mining telah diimplementasikan kedalam berbagai bidang, diantaranya dalam bidang bisnis atau perdagangan, bidang pendidikan, dan telekomunikasi.

Menurut Stanton (2013:173) data mining adalah bidang penelitian dan praktik yang berfokus pada penemuan pola-pola baru dalam data yang mengacu pada penggunaan algoritma dan komputer untuk menemukan pola baru dan menarik dalam data.

Menurut Pramudiono dalam Baskoro, dkk (2013:42) data mining adalah analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaanya.

Menurut Suyatno (2017:2) data mining ditujukan untuk mengekstrak pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basis data dan manajemen data, prapemrosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pascapemrosesan terhadap struktur yang ditemukan, visualisasi dan online updating.

Sebagai teknologi umum, data mining dapat diterapkan ke semua jenis data selama data bermakna untuk aplikasi target. Bentuk data paling dasar untuk penambangan aplikasi adalah database, data warehouse dan data transaksional. Data mining juga dapat diterapkan ke bentuk data lain (misalnya, aliran data, data urutan / urutan, grafik atau data jaringan, data spasial, data teks, data multimedia).



2.2.2 Fungsi Data Mining

Secara umum, kegunaan data mining terbagi menjadi dua yaitu deskriptif dan prediktif. Deskriptif memiliki arti untuk mencari pola-pola yang dapat dipahami manusia yang menjelaskan karakteristik data sedangkan prediktif digunakan untuk membentuk sebuah model pengetahuan guna melakukan prediksi. Berdasarkan fungsionalitasnya, tugas-tugas data mining bisa dikelompokkan menjadi enam kelompok yaitu :



Adapun penjelasan rinci dari enam kelompok tersebut sebagai berikut:

1. Klasifikasi (classification)
Proses generalisasi struktur yang diketahui untuk diaplikasikan pada data-data baru.
2. Klasterisasi (clustering)
Mengelompokkan data yang belum diketahui label kelasnya ke dalam sejumlah kelompok tertentu sesuai dengan ukuran kemiripannya.
3. Regresi (regression)
Menemukan suatu fungsi yang memodelkan data dengan kesalahan prediksi seminimal mungkin.
4. Deteksi anomali (anomaly detection)
Mengidentifikasi data yang tidak umum, berupa outlier (pencilan), perubahan atau deviasi yang mungkin sangat penting dan perlu investigasi lebih lanjut.
5. Pemodelan kebergantungan (Dependency modeling)

Mencari relasi antar tabel.

6. Perangkuman (summarization)

Menyediakan representasi data yang lebih sederhana, meliputi visualisasi dan pembuatan laporan.

2.2.3 Teknik Pembelajaran Data Mining

Teknik yang digunakan dalam data mining erat kaitannya dengan penemuan dan pembelajaran yang terbagi dalam tiga metode utama pembelajaran yaitu :

1. *Supervised learning*

Teknik yang melibatkan fase pelatihan dimana data pelatihan historis yang karakter-karakternya dipetakan ke hasil-hasil yang telah diketahui dan diolah dalam algoritma data mining. Proses ini melatih algoritma untuk mengenali variabel-variabel dan nilai-nilai kunci yang nantinya akan digunakan sebagai dasar dalam membuat perkiraan-perkiraan ketika diberikan data baru.

2. *Unsupervised learning*

Teknik pembelajaran yang tidak melibatkan fase pelatihan seperti supervised learning yakni bergantung pada penggunaan algoritma yang mendeteksi semua pola yang muncul dari kriteria penting yang spesifik dalam data masukan. Pendekatan ini mengarah pada pembuatan banyak aturan yang mengkarakteristikan penemuan associations, clusters dan segment yang kemudian dianalisis untuk menemukan hal-hal yang penting.

3. *Reinforcement learning*

Teknik yang memiliki penerapan-penerapan yang terus dioptimalkan dari waktu ke waktu dan memiliki kontrol adaptif. Menyerupai kehidupan nyata yaitu seperti "on job training" dimana seorang pekerja diberikan sekumpulan tugas yang membutuhkan keputusan-keputusan. Reinforcement learning sangat tepat digunakan untuk menyelesaikan masalah-masalah sulit yang bergantung pada waktu.

2.2.4 Proses Data Mining

Data mining biasanya terdiri dari empat proses (Stanton 2013:173) :

1. Persiapan data

Melibatkan memastikan bahwa data diatur dengan cara yang benar, bahwa bidang data yang hilang terisi, bahwa data yang tidak akurat berada dan diperbaiki atau dihapus, dan data tersebut "didaur ulang" seperlunya.

2. Analisis data eksploratori

Proses eksplorasi juga melibatkan mencari keluar nilai-nilai yang tepat untuk parameter kunci.

3. Pengembangan model

Yaitu menguji pilihan penambangan data yang paling sesuai teknik. Tergantung pada struktur dataset dan memilih yang paling menjanjikan di dalamnya sebagai sains.

4. Interpretasi hasil.

Berfokus untuk memahami dari apa algoritma data mining telah dihasilkan yang merupakan langkah penting dari perspektif pengguna data, karena ini adalah tempat kesimpulan yang dapat ditindaklanjuti terbentuk.

Beberapa tahun terakhir data tumbuh menjadi semakin heterogen dan kompleks dengan volume yang meningkat cepat secara eksponensial. Selain itu, beberapa faktor pendorong kemajuan yang berlanjut dalam bidang data mining ialah:

1. Pertumbuhan yang cepat dalam pertumbuhan data.
2. Penyimpanan data dalam data warehouse, sehingga seluruh perusahaan memiliki akses ke dalam database yang handal.
3. Adanya peningkatan akses data melalui navigasi web dan internet.
4. Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi).
5. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

2.2.5 Pengelompokan Teknik *Data Mining*

Menurut Baskoro,dkk (2013:43) data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. Classification

Suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan.

2. Association

Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian.

3. Clustering

Digunakan untuk menganalisis pengelompokan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokan belum didefinisikan sebelum dijalankannya tool data mining.

4. Forecasting

Teknik forecasting sebagai input kemudian akan mengambil sederetan angka yang menunjukkan nilai yang berjalan seiring waktu dan kemudian Teknik forecasting ini akan menghubungkan nilai masa depan dengan menggunakan bermacam-macam teknik machine learning dan teknik statistik yang berhubungan dengan musim, trend, dan noise pada data.

5. Prediction

Prediction (prediksi), untuk memperkirakan nilai masa mendatang, misalnya memprediksi stok barang satu tahun ke depan.

2.3 K-means

2.3.1 Pengertian *k-means*

K-means merupakan algoritma klasterisasi yang paling tua dan paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah karena kemudahan implementasinya. Menurut Suyanto (2017:262) Algoritma *k-means* bekerja dengan empat langkah, yaitu :

1. Himpunan data yang akan diklasterisasi, dipilih sejumlah k objek secara acak sebagai *centroid* awal.
2. Setiap objek yang bukan *centroid* dimasukkan ke *cluster* terdekat berdasarkan ukuran jarak tertentu.
3. Setiap *centroid* diperbarui berdasarkan rata-rata dari objek yang ada di dalam setiap *cluster*.
4. Langkah kedua dan ketiga dilakukan secara diulang-ulang (diiterasi) sampai semua *centroid* stabil dalam arti semua *centroid* yang dihasilkan dalam iterasi saat ini sama dengan semua *centroid* yang dihasilkan pada iterasi sebelumnya.

Berikut ini adalah langkah-langkah algoritma *k-means* :

1. Penentuan cluster awal

Dalam menentukan n buah pusat cluster awal dilakukan pembangkitan bilangan random yang merepresentasikan urutan data input. Pusat awal cluster didapatkan dari data sendiri bukan dengan menentukan titik baru, yaitu dengan random pusat awal dari data.

2. Perhitungan jarak dengan pusat cluster

Untuk mengukur jarak antar data dengan pusat dengan cluster digunakan euclidian distance, algoritma perhitungan jarak data dengan pusat cluster :

- a. Pilih nilai data dan nilai pusat cluster
- b. Hitung euclidian distance data dengan tiap pusat cluster

$$d(x_i, \mu_j) = \sqrt{(x_i - \mu_j)^2} \dots (1)$$

Penjelasan :

x_i : Data kriteria

μ_j : Centroid pada cluster ke j

3. Pengelompokan data

Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat cluster, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat cluster terdekat.

Adapun cara pengelompokan data tersebut adalah :

- a. Pilih nilai jarak tiap pusat cluster dengan data.
 - b. Cari nilai jarak terkecil.
 - c. Kelompokkan data dengan pusat cluster yang memiliki jarak terkecil.
- ## 4. Penentuan pusat cluster baru

Untuk mendapatkan pusat cluster baru bisa dihitung dari rata-rata nilai anggota cluster dan pusat cluster. Pusat cluster yang baru digunakan untuk melakukan iterasi selanjutnya, jika hasil yang didapatkan belum konvergen. Proses iterasi akan berhenti jika telah memenuhi maksimum iterasi yang dimasukkan oleh user atau hasil yang dicapai sudah konvergen (pusat cluster baru sama dengan pusat cluster lama).

Algoritma penentuan pusat cluster :

- a. Cari jumlah anggota tiap cluster
- b. Hitung pusat baru dengan rumus

$$\mu_j(t+1) = \frac{1}{N_{sj}} \sum_{x_j \in S_j} x_j \dots (2)$$

Penjelasan :

$\mu_j(t+1)$: Centroid baru pada iterasi ke 1

N_{sj} : Banyak data pada cluster s_j

Hasil dari operasi clustering yang terbentuk selanjutnya akan di evaluasi menggunakan Davies bouldin index yang dihitung dengan persamaan :

$$DBI = \frac{1}{K} \sum R_i$$

$$R_i = \max_{j=1 \dots k, i \neq j} R_{ij}$$

$$R_{ij} = \frac{\text{var}(C_i) + \text{var}(C_j)}{\|c_i - c_j\|}$$

Dimana:

C_i = Cluster i dan c_i adalah *centroid* dari cluster i

2.3.2 Keuntungan dan Kekurangan *k-means*

Sebagai fungsi penambangan data, analisis *cluster* dapat digunakan sebagai alat yang berdiri sendiri untuk memperoleh wawasan ke dalam distribusi data. Adapun keuntungan lain dari metode ini (Han,dkk , 2012:445) antara lain :

1. K-means juga disebut segmentasi data di beberapa aplikasi karena pengelompokan mempartisi set data besar ke dalam grup sesuai dengan kemiripannya.
2. K-means bisa juga digunakan untuk deteksi outlier (nilai yang "jauh" dari mana pun cluster).
3. K-means mempartisi sekumpulan objek data (atau pengamatan) ke dalam himpunan bagian, sehingga banyak digunakan dalam banyak aplikasi seperti intelijen bisnis, pengenalan pola gambar, pencarian web, biologi, dan keamanan.

Selain itu, metode clustering memiliki beberapa kekurangan (Suyanto, 2017:262) antara lain :

1. *K-means* tidak dapat menjamin konvergen pada optimum global.
2. *K-means* sering terjebak pada optimum lokal, dimana centroid akhir yang dihasilkan tidak benar-benar menjadi pusat *cluster* yang sesungguhnya.
3. Keluaran dari *k-means* bergantung pada *centroid* awal yang ditentukan secara acak.

2.4 Rapidminer

Menurut Baskoro,dkk (2013:8) Rapidminer merupakan perangkat lunak yang bersifat terbuka (open source). Rapidminer adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. Rapidminer menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik.

Rapidminer memiliki beberapa sifat sebagai berikut:

1. Ditulis dengan bahasa pemrograman *java* sehingga dapat dijalankan di berbagai sistem operasi.
2. Proses penemuan pengetahuan dimodelkan sebagai operator *trees*.
3. Representasi XML internal untuk memastikan format standar pertukaran data.
4. Bahasa *scripting* memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
5. Konsep *multi-layer* untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
6. Memiliki GUI, command line mode, dan *java* API yang dapat dipanggil dari program lain.

BAB 3

METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian adalah suatu tempat yang akan diselidiki dalam kegiatan penelitian untuk menelusuri masalah dan menerapkan hasil dari penelitian tersebut. Penelitian ini dilakukan di rumah masing-masing dimulai dari tanggal 30 Desember 2021.

3.2 Jenis dan Sumber Data

3.2.1 Jenis Data

Dalam penelitian ini kami menggunakan jenis data kuantitatif yang dijadikan sebagai pendukung dalam penyelesaian tugas ini. Definisi dan Jenis dari data yang di ambil oleh penulis dari objek penelitian yaitu menggunakan Data Kuantitatif. Data kuantitatif adalah data dari hasil penelitian yang bersifat terstruktur atau berpola sehingga ragam data yang diperoleh dari sumber riset lebih mudah dibaca oleh peneliti.

3.3 Sumber Data

Sumber data yang digunakan penulis dalam mendukung penelitian untuk menyelesaikan tugas akhir ini yaitu data primer dan data sekunder. Adapun pengertian dan contoh dari data yang diambil penulis pada objek penelitian adalah:

1) Data Primer

Data primer adalah jenis data yang dikumpulkan secara langsung dari sumber utamanya seperti melalui wawancara, survei, dataset statistik, dan sebagainya. Dalam pengumpulan data primer dalam penelitian ini menggunakan metode dataset statistik yang dimana penggunaan dataset statistik ini merupakan penggunaan data yang sudah tersedia.

2) Data Sekunder

Data sekunder adalah data pendukung yang sumbernya didapat dari sumber yang telah ada atau peneliti sebagai tangan kedua. Data sekunder dapat diperoleh dari berbagai sumber seperti laporan, jurnal, dan lainya. Data sekunder yang

digunakan dalam penelitian ini adalah data yang berhubungan dengan data sebelumnya.

BAB 4

PEMBAHASAN

4.1 Proses *clustering*

Pada tahap ini akan dilakukan proses utama yaitu segmentasi atau pengelompokan data angka garis kemiskinan. Berikut ini merupakan penerapan algoritma *k-means* dengan asumsi bahwa parameter *input* adalah jumlah dataset sebanyak n data dan jumlah inisialisasi *centroid* $k = 3$ sesuai dengan penelitian. Data yang diambil untuk penelitian berjumlah 34 untuk dijadikan contoh penerapan algoritma *k-means*. Percobaan dilakukan dengan menggunakan parameter-parameter berikut :

Jumlah *cluster* : 4

Jumlah data : 107

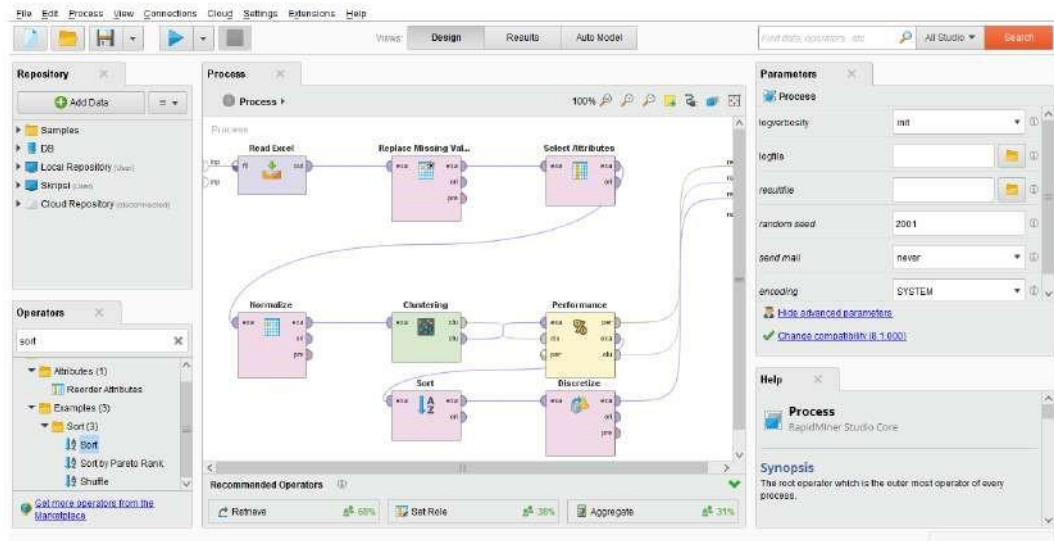
Jumlah atribut : 7

4.2 Pengujian *rapidminer*

Pada penelitian ini penulis menggunakan *tool rapidminer* sebagai alat pengujian dataset.

Adapun tahapan pengujian yang dilakukan yaitu sebagai

berikut :



Pada tahapan ini dilakukan 8 proses yaitu :

a. *Read excel*

Tahapan ini dilakukan operasi penginputan dataset berupa file berekstensi .xls angka garis kemiskinan di Indonesia.

b. *Replace missing value*

Tahapan ini dilakukan operasi pengisian nilai yang hilang dengan nilai maksimal.

c. *Select attributes*

Tahapan ini dilakukan operasi pemilihan atribut yang akan dihitung yaitu data perdesaan dan data perkotaan pada bulan September 2017.

d. *Normalize*

Tahapan ini dilakukan operasi normalisasi data menggunakan metode *z score* dihasilkan nilai-nilai yang sudah distandarkan.

e. *Clustering*

Tahapan ini dilakukan operasi *clustering* sebagai algoritma yang digunakan pada penelitian ini.

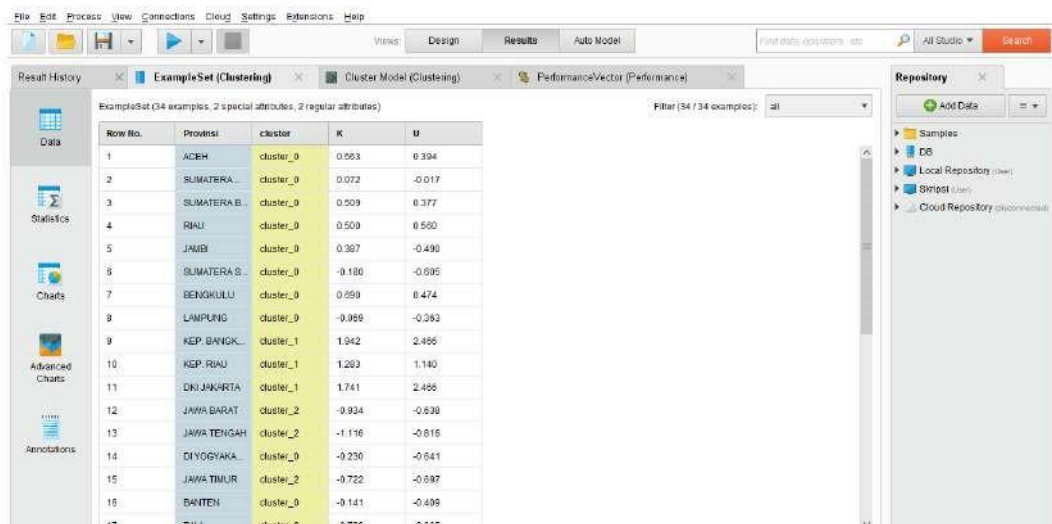
f. *Performance*

Tahapan ini dilakukan operasi pencarian nilai *davies bouldin index*. g. Sort

Tahapan ini dilakukan operasi pengurutan anggota *cluster* 0 sampai dengan *cluster* 2.

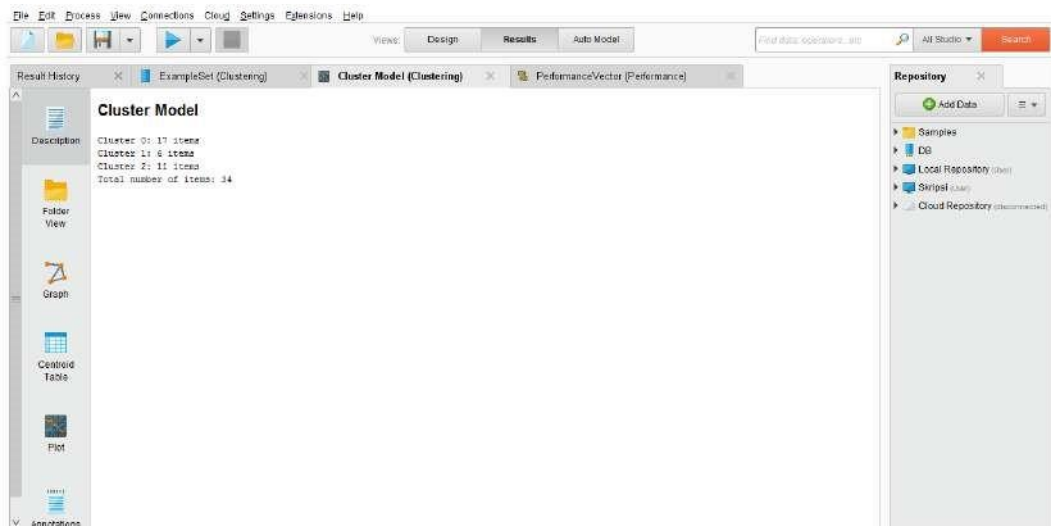
h. Discretize

Tahapan ini dilakukan operasi mengubah atribut numerik yang dipilih menjadi atribut nominal dengan mendiskritisasi atribut numerik.



Row No.	Provinsi	cluster	K	U
1	ACEH	cluster_0	0.053	0.394
2	SUMATERA	cluster_0	0.072	-0.017
3	SUMATERA B.	cluster_0	0.509	0.377
4	RIAU	cluster_0	0.509	0.550
5	JAWA	cluster_0	0.367	-0.490
6	SUMATERA B.	cluster_0	-0.186	-0.095
7	BENGKULU	cluster_0	0.059	0.474
8	LAMPUNG	cluster_0	-0.956	-0.282
9	KEP. BANGK.	cluster_1	1.942	2.456
10	KEP. RIAU	cluster_1	1.283	1.140
11	DKI JAKARTA	cluster_1	1.741	2.456
12	JAWA BARAT	cluster_2	-0.934	-0.630
13	JAWA TENGAH	cluster_2	-1.116	-0.615
14	DIYOGYAKA	cluster_0	-0.230	-0.641
15	JAWA TIMUR	cluster_2	-0.722	-0.687
16	BAWITEN	cluster_0	-0.141	-0.406
17	RIAU	cluster_0	-0.716	-0.585

Pada tahapan ini ditampilkan hasil dari klasterisasi data. Label *cluster* terbagi menjadi tiga kelompok yaitu *cluster* 0, *cluster* 1, *cluster* 2. Pembagian ini berdasarkan hasil kedekatan tiap masing-masing data dengan jarak terdekat (k).



Pada tahapan ini ditampilkan hasil pembagian data terhadap tiap *cluster*.

Cluster 0 memiliki 17 anggota, *Cluster* 1 memiliki 6 anggota, *Cluster* 2 memiliki 11 anggota dari total 34 dataset yang di uji.

//Local Repository/Clustering/Graph* - RapidMiner Studio Educational 9.10.001 @ AbigailPerkasa

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators... etc All Studio

ExampleSet (/Local Repository/Clustering/Data Set Kelompok 3 Jam 1) ExampleSet (/Local Repository/Clustering/Data Set Kelompok 3 Jam 1)

PerformanceVector (Performance) ExampleSet (/Local Repository/Kelompok 3)

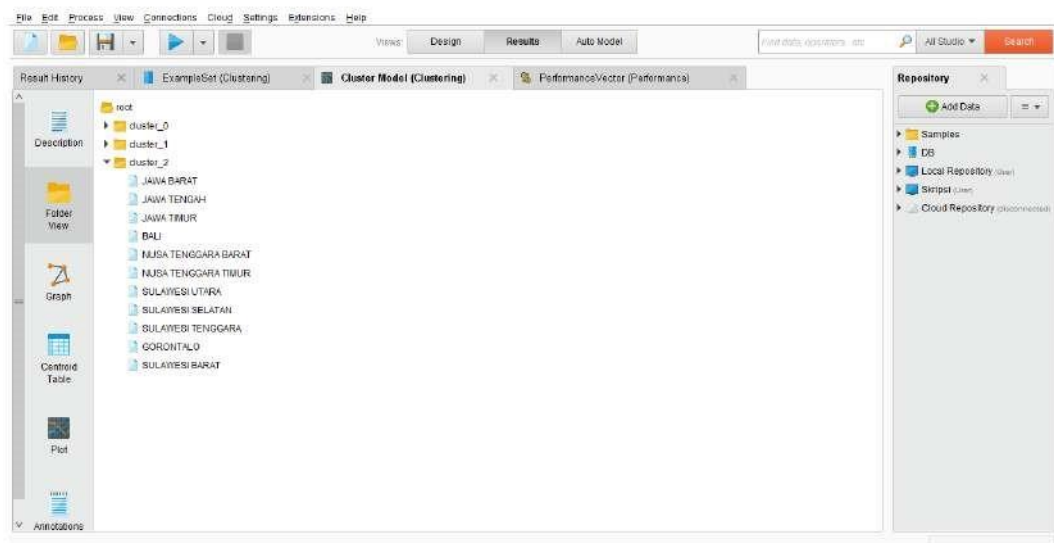
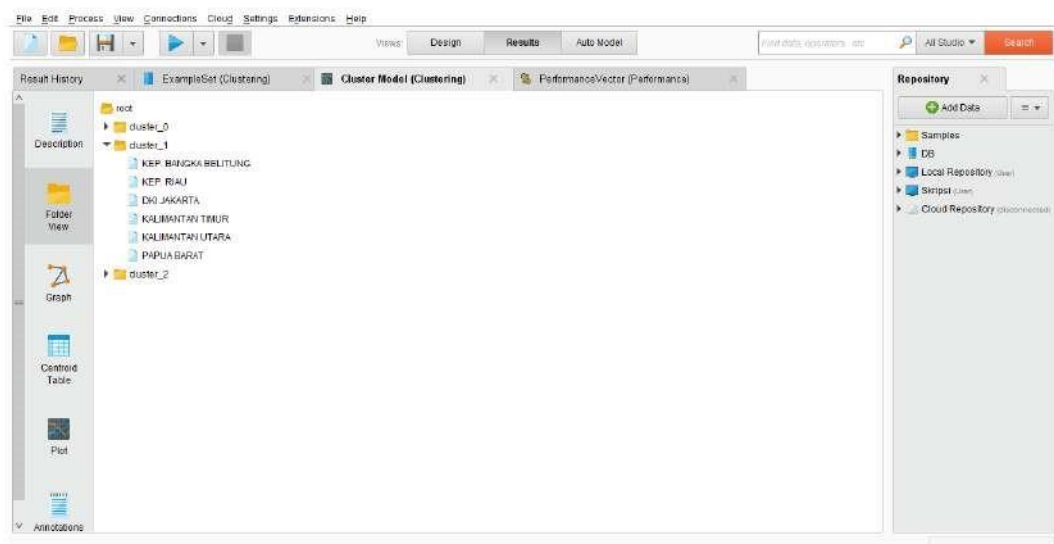
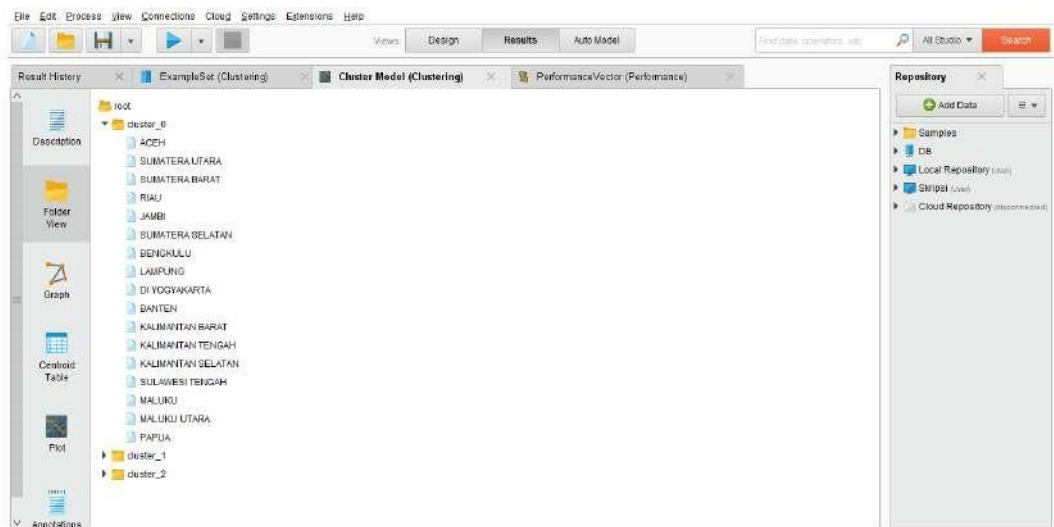
PerformanceVector (Performance (4)) PerformanceVector (Performance (3)) PerformanceVector (Performance (2))

Result History Cluster Model (Clustering (2)) ExampleSet (Clustering (2)) PerformanceVector (Performance (5))

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Fisika	72.833	81.435	79.789	92.652	74.588
Biologi	88.750	79.652	80.947	81.304	73.529
Matematika	84.375	92.261	70.158	80.696	80.059
Kimia	88.917	74.348	79	90	72.176
Bahasa Inggris	84.458	72.826	73.684	84.087	87.118

11:57 AM 1/6/2022

Pada tahapan ini ditampilkan nilai titik pusat pada tiap *cluster*. Nilai tersebut menjadikan acuan perhitungan pada tiap-tiap dataset dengan cara mengukur kedekatan nilai dengan masing-masing titik pusat *cluster*.

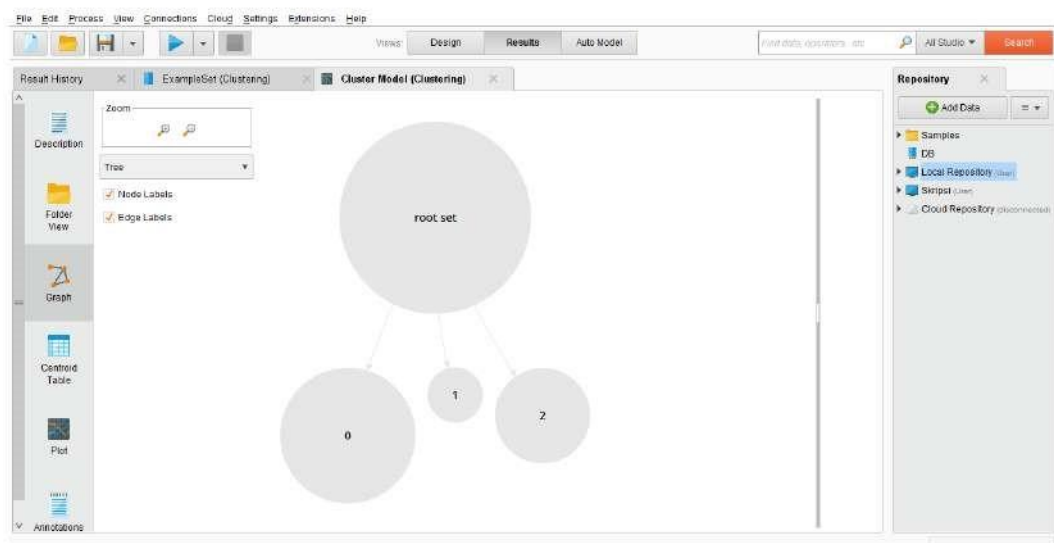


Gambar 4. 7 Anggota cluster 2

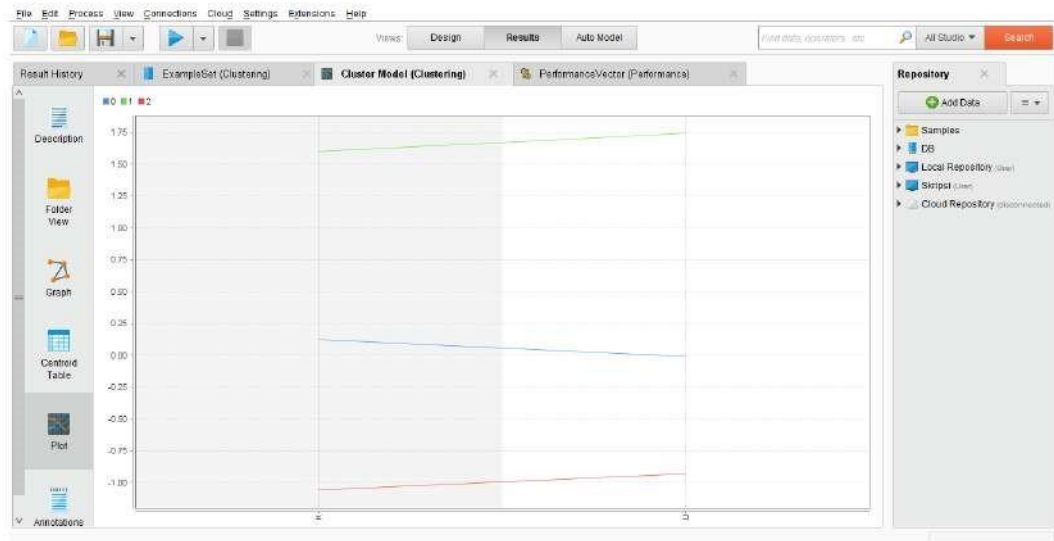
Name	Type	Missing	Statistics	Filter (4 / 4 attributes)
Provinsi	Polynomial	0	Min: SUMATERA UTARA (1) Max: ACEH (1)	Value: ACEH (1), BALI (1), ... (32 n)
cluster	Nominal	0	Min: cluster_1 (6) Max: cluster_0 (17)	Value: cluster_0 (17), cluster_2 (1)
K	Real	0	Min: -1.545 Max: 1.951	Average: -0.000
U	Real	0	Min: -1.380 Max: 2.466	Average: -0

Showing attributes 1 - 4 Examples: 34 Special Attributes: 2 Regular Attributes: 2

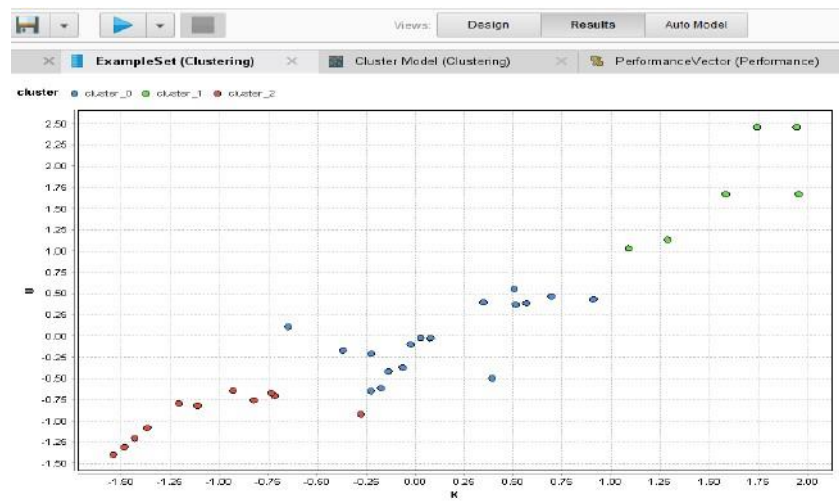
Pada tahapan ini ditampilkan hasil statistik dari data yang sudah di uji. Pada tabel K dan U terdapat 3 atribut yaitu *min* sebagai nilai terendah pada tabel dataset, *max* sebagai nilai tertinggi pada tabel dataset dan *average* sebagai nilai rata-rata dari penjumlahan tabel dataset tersebut.



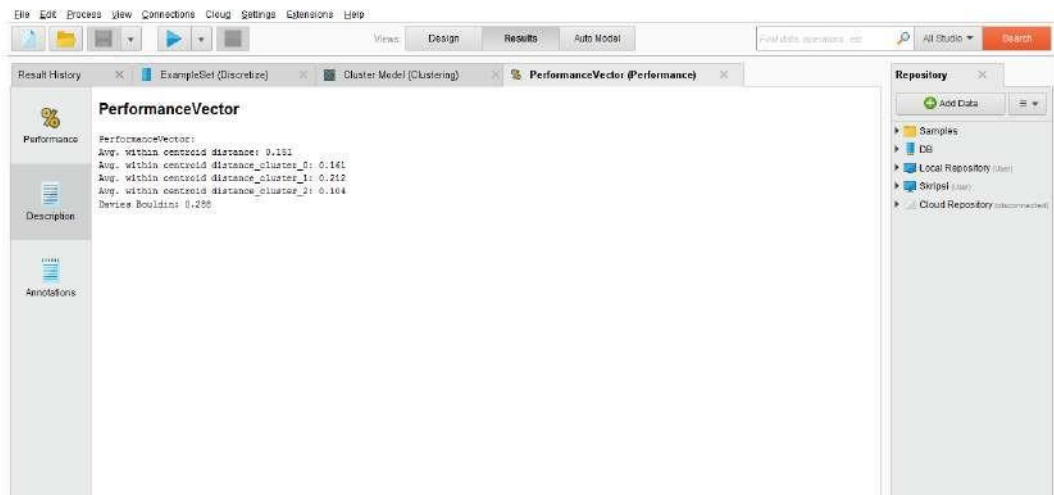
Pada tahapan ini ditampilkan hasil pembagian dari tiga kelompok berupa lingkaran. Ukuran tiap lingkaran mendeskripsikan jumlah banyaknya anggota tiap *cluster*.



Pada tahapan ini ditampilkan hasil plot dari hasil pengujian. Pada bagian sebelah kiri menunjukkan angka pedapatan dan bagian sebelah kanan ditampilkan garis sebagai gambaran rataaan nilai anggota.

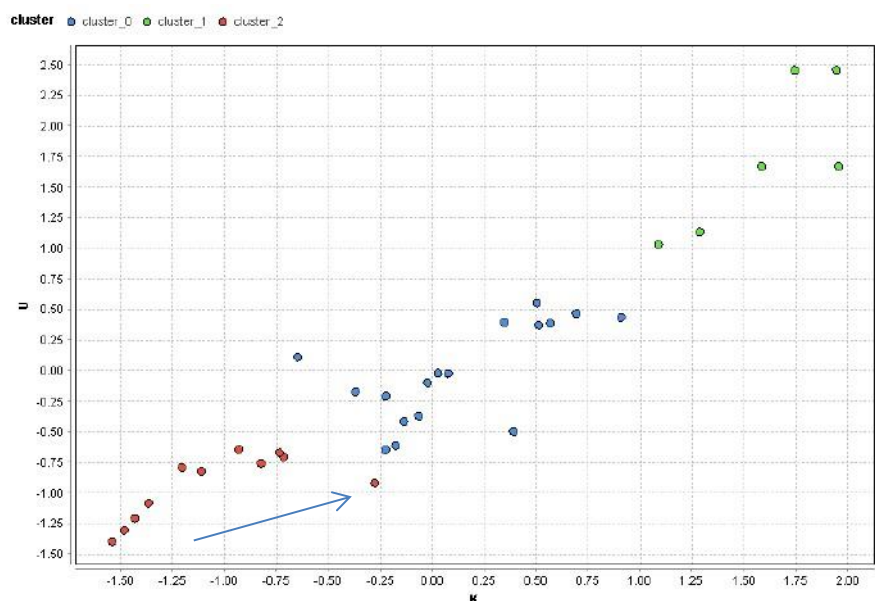


Pada tahapan ini ditampilkan hasil pengelompokan data dalam bentuk grafik titik dengan warna. Warna biru mengartikan *cluster* 0, warna hijau mengartikan *cluster* 1 dan warna merah mengartikan *cluster* 2.



4.3 Pembahasan hasil *clustering k-means*

Setelah dilakukan pengujian dengan *tool rapidminer*, maka didapatkan kesimpulan sebagai berikut :



Tanda panah biru adalah data pada provinsi Nusa Tenggara Timur yang memiliki kedekatan jarak secara sekilas dekat dengan *cluster* 0 yang diartikan dengan titik berwarna biru. Peneliti berpendapat bahwa data tersebut tidak dapat bergabung dengan *cluster* 0 dikarenakan nilai z

score menunjukkan hasil yang lebih dekat dengan titik pusat *cluster* 2 sehingga data tersebut dinyatakan sebagai anggota *cluster* 2.

Selanjutnya dibuat himpunan dan domain untuk masing-masing variabel untuk memudahkan mendeskripsikan tiap provinsi:

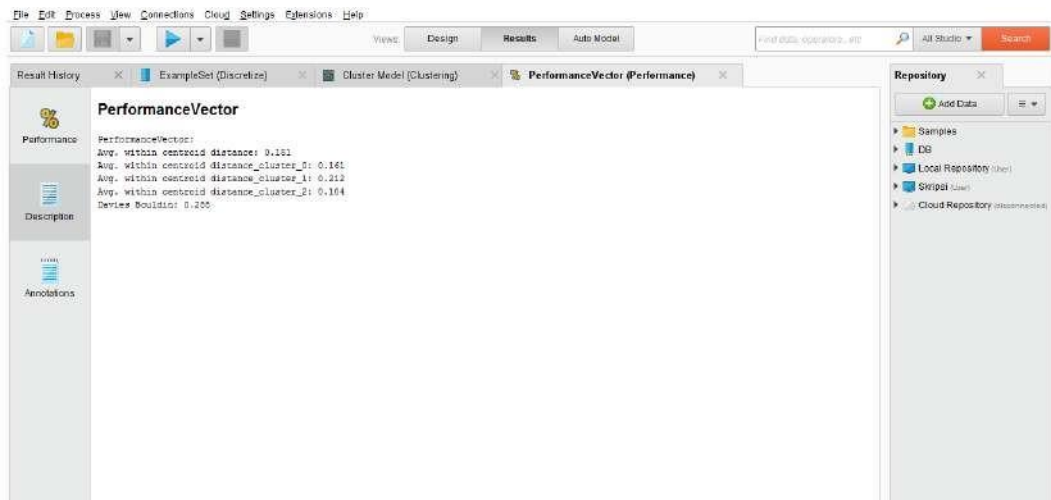
Row No.	Provinsi	cluster	K	U
13	KALIMANTAN...	cluster_0	[413714.0 - 5...	[369537.5 - 4...
14	SULAWESI T...	cluster_0	[413714.0 - 5...	[369537.5 - 4...
15	MALUKU	cluster_0	[413714.0 - 5...	[369537.5 - 4...
16	MALUKU LITA...	cluster_0	[413714.0 - 5...	[369537.5 - 4...
17	PAPUA	cluster_0	[413714.0 - 5...	[369537.5 - 4...
18	KEP. BANGK...	cluster_1	[515982.0 - ∞)	[478227.0 - ∞)
19	KEP. RIAU	cluster_1	[515982.0 - ∞)	[478227.0 - ∞)
20	DKI JAKARTA	cluster_1	[515982.0 - ∞)	[478227.0 - ∞)
21	KALIMANTAN...	cluster_1	[515982.0 - ∞)	[478227.0 - ∞)
22	KALIMANTAN...	cluster_1	[515982.0 - ∞)	[478227.0 - ∞)
23	PAPUA BARAT	cluster_1	[515982.0 - ∞)	[478227.0 - ∞)
24	JAWA BARAT	cluster_2	[∞ - 413714.0]	[∞ - 369537.5]
25	JAWA TENGAH	cluster_2	[∞ - 413714.0]	[∞ - 369537.5]
26	JAWA TIMUR	cluster_2	[∞ - 413714.0]	[∞ - 369537.5]
27	BALI	cluster_2	[∞ - 413714.0]	[∞ - 369537.5]
28	NUSA TENG...	cluster_2	[∞ - 413714.0]	[∞ - 369537.5]

Setelah *cluster* terbentuk, maka dapat diperoleh kesimpulan sebagai berikut:

Tabel 4. 7 Deskripsi data berdasarkan dengan *cluster*

Cluster	Deskripsi
0	Anggota <i>cluster</i> dengan rata-rata penghasilan rendah
1	Anggota <i>cluster</i> dengan rata-rata penghasilan cukup
2	Anggota <i>cluster</i> dengan rata-rata penghasilan sangat rendah

Setelah proses *clustering* selesai selanjutnya dilakukan operasi *performance* untuk mengetahui nilai dari *Davies bouldin index* yang bertujuan untuk memaksimalkan pengukuran jarak antar *cluster* dan meminimalkan jarak antar anggota dalam suatu *cluster* .



Hasil yang diperoleh dari operasi *performance vector* adalah sebagai berikut:

Tabel 4. 8 Hasil operasi *performance vector*

<i>Performance Vector</i>	<i>Value</i>
<i>Avg. within centroid distance</i>	0.151
<i>Avg. within centroid distance cluster 0</i>	0.161
<i>Avg. within centroid distance cluster 1</i>	0.212
<i>Avg. within centroid distance cluster 2</i>	0.104
<i>Davies Bouldin</i>	0.288

Evaluasi hasil dari *average within centroid distance* mendekati angka 0 mengartikan bahwa masing-masing anggota didalam *cluster* berada dalam jarak yang berdekatan. Evaluasi menggunakan *davies bouldin index* memiliki skema internal *cluster* yang dilihat dari kuantitas dan kedekatan antar hasil *cluster*. Semakin kecil nilai *davies bouldin index* yang diperoleh (non-negatif ≥ 0), maka semakin baik *cluster* yang diperoleh dari pengelompokan menggunakan metode *clustering*. Hasil perhitungan menggunakan algoritma *k-means* menunjukan nilai 0,288. Angka tersebut memiliki arti masing-masing objek dalam cluster tersebut memiliki kesamaan yang cukup baik karena mendekati angka 0.

BAB 5

PENUTUP

1.1 Kesimpulan

Berdasarkan hasil penelitian yang dilakukan oleh penulis, dapat diambil kesimpulan untuk setiap cluster dari data set memiliki nilai lebih dari 72 yang berarti setiap siswa memiliki nilai diatas rata-rata.

1.2 Saran

Saat pemilihan data set usahakan untuk data atribut lebih dari 100 an agar mudah menemukan algoritma clusteringnya.

DAFTAR PUSTAKA

Aprilla, D., Ambarwati, L., Baskoro, D. A., Wicaksana, I. W. S. 2013. Belajar Data Mining dengan RapidMiner. Jakarta: Open Content Model

Aziz, A., Purmaningsih, C., Saptono, R. 2014. Pemanfaatan Metode K-means Clustering Dalam Penentuan Penjurusan Siswa SMA. Jurnal ITSMART. Vol 3 (1): 27-33

Hamzah, A., Syechalad, M. N., Takdir, A. 2013. Analisis Kemiskinan Rumah Tangga Berdasarkan Karakteristik Sosial Ekonomi Di Kabupaten Aceh Barat Daya. Vol. 1. Page 67-75

Han, J., Kamber, M., Pei, J. 2012. Data Mining Concepts and Techniques. Waltham: Elsevier.

Jumadi, B.D.S. 2018. Peningkatan Hasil Evaluasi Clustering Davies Bouldin Index Dengan Penentuan Titik Pusat Cluster Awal Algoritma K-means [skripsi]. Medan. Universitas Sumatera Utara

Muhidin ,A. 2017. Analisa Metode Hierarchical dan K-means Dengan Model LRFMP Pada Segmentasi Pelanggan. SIGMA. Vol 7 (1): 81-88

LAMPIRAN

Lampiran 1. Tabel Data set Penelitian

NIS	Nama	Fisika	Biologi	Matematika	Kimia	Bahasa Inggris
2012173	ABDUL JABAR	80	86	83	76	73
2012174	ADITIA ALFARISI	66	69	89	65	95
2012175	ADITYA PRAMUDITA	74	74	94	84	94
2012176	AINUN ZARIYAH	95	96	82	79	77
2012177	ALIEF NOVALYANSYAH	71	66	93	67	88
2012178	ALIEFVIA REZQA	77	76	69	74	93
2012179	ALMA SHARIKA SOFYANTI	74	87	77	86	79
2012180	ALVIEN ALFARIZI SANTOSO	82	81	80	72	96
2012181	AMARULAH ABDUL HAMID	87	83	88	72	67
2012182	ARI HERMANSYAH	94	75	68	94	89
2012183	AYU DWI CAHYANI	83	80	78	85	80
2012184	DIAN NATULHIKMAH	92	89	74	84	70
2012185	DIAN SUNARSIH	68	86	70	72	66
2012186	DINI WIDIYA OKTAPIANI	86	69	93	82	94
2012187	DWI SAKINA	81	90	86	95	65
2012188	EKA SUPARTINI	68	72	89	97	97
2012189	FAHRIYAN HARIS	90	74	74	90	96
2012190	FITRI DIANSARI	67	71	76	65	94
2012191	FITRIANI	76	92	77	72	69
2012192	HANISYA RAHMI NOVIA SUMBAWATI	80	66	96	91	97
2012193	IIN PUTRI ANDANI	67	95	83	81	67
2012194	IKHSAN SAPUTRA	77	74	84	66	91
2012195	ILA MULYANI	73	94	90	86	65
2012196	INDA SARI	81	89	68	71	87
2012197	INDAS KHOFIFAH	92	84	91	71	79
2012198	JULIANA FEBRIANI	65	92	76	85	68
2012199	KHAERUL ANNAM	74	97	97	74	76
2012200	LALA APRILIA SALSABILA	79	92	81	90	84
2012201	LELY RAHMAWATI	68	83	81	73	97
2012202	M REZA JULIANTO	89	74	86	74	87
2012203	MINI SEPTIKA	81	79	90	66	70
2012204	MOHAMMED SESAY	90	68	73	71	91
2012205	MUHAMAD IKSAN	83	78	92	89	84
2012206	MUHAMMAD IQBAL FUSTHAHULLAH	88	66	69	76	86

2012207	MUHAMMAD RAMDANI	76	88	93	70	90
2012208	MUHAMMAD WAHYU MIFTAH JUANDI	76	74	82	73	67
2012209	NADIA	81	97	85	92	89
2012210	NADIA SAFITRI	74	76	68	78	78
2012211	NANDA LARA SAFITRI	89	85	93	75	83
2012212	NAWAB WINARDI	78	94	66	83	88
2012213	NENGSIH LESTARI	88	90	70	75	86
2012214	NUNUNG AFRIANI	74	97	85	73	91
2012215	NURILA	88	75	89	75	89
2012216	NURWAHIDA AFLIA	67	73	75	66	73
2012217	PRASTIA JULIA UTAMI	85	68	89	97	79
2012218	RAMONA DWI PUTRI	77	74	73	95	66
2012219	RATU WIDIA	77	86	67	68	92
2012220	RENI OKTAPIANI	67	71	82	80	74
2012221	RIRIN TRIANI	79	70	78	68	73
2012222	SARTIKA OKTARIDA	71	76	74	90	79
2012223	SATRIA HERNAWAN	80	70	88	67	75
2012224	SATRIA MOKTAR	84	80	78	77	91
2012225	SEPTIANA EKA PRATIWI/putri	95	66	86	86	90
2012226	SITI AMNAL ASKIYA	75	81	94	91	91
2012227	SITI RAHMAWATI	93	91	72	85	66
2012228	SONI APRIAWAN	67	66	76	96	65
2012229	SRI RAHAYU	66	69	97	72	96
2012230	SRI WULANDARI	73	80	96	77	73
2012231	SURYA YUDIANTO	84	65	79	87	83
2012232	TIARA FITRI RAMDANI	77	85	87	80	87
2012233	TRIANA RIZTA MUHARROZI	89	75	74	85	75
2012234	VIDYANA WULANDARI	75	85	72	79	67
2012235	VINA FEBRIANA	84	65	72	66	89
2012236	WULANDARI AHDIAT	82	68	90	85	83
2012237	YOPIN INTAN SEPTIANTI	87	96	75	78	91
2012238	YULIA PUTRI ANGGINI	88	74	86	91	83
2012239	YUYUN FEBRIANTY	92	81	72	81	72

2012240	AINI FEBRIANTI	83	71	95	81	91
2012241	ALICHA SYEHAN	81	70	65	65	97
2012242	ALVINA SUMANTY	93	87	97	92	71
2012243	ANA SAPITRI	83	67	95	83	72
2012244	ANDI RAHMAT HIDAYAT	65	68	79	82	89
2012245	ANIS SELPIANI	72	89	81	93	78
2012246	ANJAS MARA TRI SYAPUTRA	95	66	73	97	95
2012247	ANUGRAH ADE CANTARY	96	87	71	93	73
2012248	ARDIANSYAH	87	83	69	86	90
2012249	ARMELIA PUSPITASARI	89	97	92	94	89
2012250	ASRI ARSITA	96	72	69	83	73
2012251	AZRIL NURAQSYA PRANA	76	69	78	86	66
2012252	CANRIKA SYAHPUTRI	83	83	91	97	87
2012253	DEBBY TRI CAHYANI	75	74	68	71	76
2012254	DEWI LESTARI	67	79	92	70	91
2012255	DISA SELPIAH	72	82	80	69	79
2012256	DWI FEBRI AMANDARI	82	85	87	68	73
2012257	DWI PUTRI	85	93	92	77	71
2012258	DWI YANA RAVIKA	73	88	90	91	65
2012259	EFA ROSIFA	90	87	81	77	87
2012260	EKA RAHMAT TULYANTI	69	66	78	73	85
2012261	EKA SAPUTRIANI	72	91	70	83	75
2012262	ELI ERMAWATI	65	78	97	86	87
2012263	ERIK SOLIKIN	95	84	95	97	86
2012264	ERNA WAHYUNI	81	83	91	92	65
2012265	FADILLAH EKA MEILANY	72	79	73	75	73
2012266	FARHAN	79	71	70	94	92
2012267	FAUZI HAMDANI	81	73	91	74	93
2012268	FITRI DAMAYANTI	69	67	73	78	82
2012269	GUSTINA REJAUNA	91	81	93	69	78
2012270	HADID SURYADIN	81	71	66	68	66
2012271	HAMDANI	72	90	71	87	93

2012272	HERNI NUR AZIZA	72	66	87	85	67
2012273	HIDAYATULLAH	73	79	73	88	90
2012274	IDRIS FEBRIANSYAH	81	97	87	78	73
2012275	IKHSAN ADITYA	81	94	84	92	92
2012276	ILHAM	68	95	67	83	74
2012277	ILHAM ANUGRAH PUTRA	70	96	88	79	89
2012278	NADYA AULIA ISNAINI	97	67	73	93	95