KLATERISASI DATA PENDUDUK BERDASARKAN PEKERJAAN MENGGUNAKAN METODE K-MEANS PADA WILAYAH JAWA BARAT

Eka Roehatul Jannah, Martanto
Manajemen Informatika, STMIK IKMI Cirebon
Jalan Perjuangan No. 10B Majasem Kota Cirebon
ekaroehatuljannah12@gmail.com

ABSTRAK

Perkembangan teknologi merupakan peluang yang tepat memperoleh data dengan lebih efektif dan efisien. Data mining adalah salah satu komponen dalam proses Knowledge Discovery in Databases (KDD). KDD adalah suatu rangkaian proses yang bertujuan menemukan informasi yang bermanfaat dari sumber data dalam database. Permasalahan dalam penelitian ini, bagaimana jika Metode K-Means mungkin tidak sesuai untuk mengelompokkan data penduduk berdasarkan pekerjaan?. Penelitian ini bertujuan untuk mengidentifikasi pola pekerjaan penduduk wilayah Jawa Barat dan membentuk kelompok pekerjaan yang serupa. Melalui metode K-Means, akan memungkinkan saya untuk mengelompokkan penduduk Jawa Barat berdasarkan jenis pekerjaan mereka menggunakan tahapan KDD. Dengan tahapan KDD kita dapat dengan mudah melihat data penduduk berdasarkan pekerjaan dari tahun 2011-2023. Dapat diambil kesimpulan bahwa penduduk yang bekerja dengan nilai tertinggi adalah pada *Cluster* 3 yang ditandai dengan warna biru (tinggi) berjumlah 151 *items*, untuk data pekerjaan dengan nilai sedang berada pada *Cluster* 2 yang ditandai dengan warna oranye (sedang) berjumlah 100 *items*, dan ntuk penjualan dengan nilai terendah yaitu pada *Cluster* 0 dan *Cluster* 1 yang ditandai dengan warna hijau dan hitam (rendah) dengan jumlah yang sama yaitu 50 *items*. Hasil percobaan yang dilakukan pada data penduduk berdasarkan pekerjaan menggunakan metode DBI (*Davies Bouldin Index*), menghasilkan nilai K terbaik pada cluster 4 yaitu 0,262.

Kata kunci : *Data Mining, Klasterisasi, K-Means*

1. PENDAHULUAN

Teknologi informasi selalu berkembang pesat. Perkembangan ini merupakan peluang yang tepat memperoleh data dengan lebih efektif dan efisien, namun dengan jenis yang berbeda. Untuk mengolah data tersebut diperlukan sesuatu teknologi yang memproses hasil atau informasi yang didapatkan ternyata cocok. Sebuah teknik yang dimanfaatkan adalah data mining. Data mining adalah media mengolah dan mengelompokkan data yang terjaring dalam suatu basis data. Basis data belum diproses dengan cara apa pun masih tradisional, karena menghasilkan informasi lebih lama dan informasi yang dihasilkan mengandung prasangka yang besar. Informasi seperti itu tidak efektif dalam mengambil keputusan.

Data mining merupakan komponen integral dari proses Knowledge Discovery in Database (KDD). Fase ini berperan dalam melakukan analisis menyeluruh terhadap pengolahan data, mengubah kumpulan data menjadi pengetahuan yang memiliki nilai dan kegunaan [1]. KDD (Knowledge Discovery in Databases) adalah suatu proses yang bertujuan untuk menemukan informasi yang berguna dari sumber data dalam database. Tahapan dalam proses KDD mencakup pemahaman terhadap bidang aplikasi yang bersangkutan, pembuatan data target yang ditentukan dari data mentah dalam database, serta melibatkan preprocessing data dan kegiatan pembersihan data [2].

Klasterisasi pada data penduduk berdasarkan pekerjaan merupakan langkah penting dalam menganalisis struktur dan karakteristik penduduk di suatu wilayah. Dalam konteks ini, Jawa Barat, sebagai salah satu provinsi terbesar di Indonesia, memiliki keragaman penduduk yang signifikan dari segi pekerjaan. Melalui metode *K-Means*, akan memumudahkan dalam mengelompokkan penduduk Jawa Barat berdasarkan jenis pekerjaan mereka. Data ini akan memberikan informasi yang berharga bagi pemerintah daerah dan pemangku kepentingan lainnya untuk merencanakan kebijakan yang lebih baik dalam bidang apapun seperti pendidikan, ketenagakerjaan, dan pemberdayaan masyarakat [3].

Pada bagian ini, saya akan menyajikan data yang mendukung pentingnya penelitian ini dilakukan, data ini di ambil dari sumber informasi mengenai data penduduk yang dikelompokkan menurut pekerjaan penduduk Provinsi Jawa Barat.

Tabel 1. Data dan Fakta

No	Data	Fakta
1	Nama Provinsi	Jawa Barat
2	Tahun Pendataan	2011-2023
3	Jumlah Penduduk	48.027.280 ±
4	Situs Web Data	https://data.jabarprov.go.id

Tabel 1 adalah sebuah tabel informasi data yang akan di gunakan penulis untuk penelitian ini. Beberapa informasi kunci yang disajikan dalam tabel tersebut meliputi nama provinsi, tahun pendataan, jumlah

penduduk, dan situs web data dimana data tersebut diambil. Data pendataan penduduk Jawa Barat dapat diakses melalui situs *web* Badan Pusat Statistik (BPS) Jawa Barat. BPS Jawa Barat menyediakan data jumlah penduduk menurut kabupaten/kota untuk tahun 2018, 2019, dan 2020,Selain itu, hasil *long-form Sensus* Penduduk 2020 Provinsi Jawa Barat juga telah dirilis pada Februari 2023. Terdapat juga data terkait pendataan penduduk *non-permanen* dan rentan administrasi kependudukan yang dapat diakses melalui situs *web* opendata.jabarprov.go.id. Berdasarkan hasil Sensus Penduduk 2020, jumlah penduduk Jawa Barat pada tahun tersebut sebanyak 48, 27 juta jiwa.

Penelitian ini dibuat karena mencerminkan fokus penelitian yang ingin mengelompokkan data penduduk berdasarkan pekerjaan menggunakan metode *K-Means*. Judul tersebut juga mencerminkan wilayah spesifik di mana penelitian ini dilakukan yaitu Jawa Barat, sehingga memberikan gambaran tentang lingkup *geografis* dari penelitian tersebut. Selain itu, pemilihan judul ini juga dapat menjadi panduan bagi pembaca untuk memahami topik utama dan pendekatan yang digunakan dalam penelitian. Hal ini dapat membantu pembaca atau peneliti lain untuk memahami konteks dan relevansi dari temuan yang akan disajikan dalam penelitian ini.

2. TINJAUAN PUSTAKA

2.1. K-Means

K-Means adalah algoritma pengelompokan populer yang digunakan dalam pembelajaran mesin dan analisis data untuk mempartisi titik-titik data ke dalam kelompok-kelompok yang berbeda atau *cluster*. Bagian ini akan membahas secara rinci aspek-aspek kunci dari metode K-Means, tujuan, kelebihan, dan keterbatasannya. Selain itu, bagian ini juga akan membahas lebih dalam tentang pentingnya mempersiapkan data sebelum menerapkan algoritma dan memilih dengan cermat jumlah cluster yang optimal berdasarkan persyaratan tertentu. Bagian ini juga akan mengeksplorasi implementasi algoritma K-Means, mengevaluasi hasilnya secara menyeluruh, dan mengatasi tantangan umum yang dihadapi selama proses pengelompokan. Selain itu, bagian ini juga akan menyajikan berbagai variasi dan perluasan K-Means, yang menampilkan kemampuan unik dan aplikasi potensial seperti segmentasi pelanggan, kompresi gambar, dan deteksi anomali.

Dengan menawarkan contoh-contoh praktis dan studi kasus, bagian yang diperluas ini bertujuan untuk memberikan pemahaman yang komprehensif tentang metode K-Means dan relevansinya dalam skenario dunia nyata. Pada akhirnya, bagian ini diakhiri dengan merangkum temuan dan implikasi yang diperoleh dari penggunaan algoritme K-Means secara ekstensif, yang selanjutnya menekankan signifikansinya sebagai alat yang berharga di bidang pembelajaran mesin dan analisis data [4]. Beberapa aspek kunci dari pengelompokan k-means meliputi:

a. Kluster : Kumpulan titik data yang dikumpulkan bersama karena kesamaan tertentu.
b. Centroid : Lokasi imajiner atau nyata yang mewakili pusat cluster.
c. Proses iteratif : Algoritme ini bergantian antara menetapkan titik data ke klaster berdasarkan titik pusat saat ini dan memilih titik pusat baru untuk iterasi berikutnya.
d. Konvergensi : Algoritma konvergen ketika centroid tidak lagi berubah, yang menunjukkan bahwa cluster telah diidentifikasi.

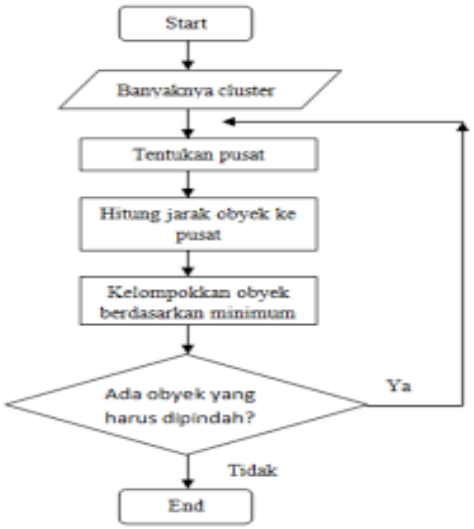
Pengelompokan K-means berguna untuk berbagai jenis data, termasuk data numerik dan kategorikal. Namun, K-means memiliki beberapa keterbatasan, seperti kebutuhan untuk menentukan jumlah *cluster* (k) sebelumnya dan sensitivitas terhadap penempatan awal *centroid*. Terlepas dari keterbatasan ini, k-means *clustering* tetap menjadi teknik yang banyak digunakan dalam pembelajaran mesin tanpa pengawasan karena kesederhanaan dan keefektifannya dalam mengelompokkan titik-titik data yang serupa. Dalam artikel yang dipublikasikan oleh Elsevier berjudul "Data Clustering: 50 Tahun Setelah K-means" [5], dijelaskan bahwa pengaturan data menjadi kluster merupakan model yang sangat dasar untuk pemahaman dan pembelajaran. Analisis kluster merupakan studi formal untuk mengelompokkan benda-benda berdasarkan karakteristik yang diukur, dengan mempertimbangkan tingkat kemiripan di antara mereka.

Clustering adalah metode pengelompokan yang menggunakan teknik unsupervised learning, di mana tidak ada fase pembelajaran yang diperlukan dan tanpa penggunaan pelabelan pada setiap kelompok. Metode pengelompokan membagi data ke dalam kelompok sehingga data yang memiliki karakteristik serupa dikelompokkan bersama dalam satu kluster yang sama [6].

Tujuan dari proses pengelompokan ini adalah untuk mengoptimalkan tujuan fungsi yang telah ditetapkan dalam pengelompokan, yang pada umumnya bertujuan untuk mengurangi variasi dalam suatu kelompok dan meningkatkan variasi antar kelompok. Pengelompokan atau analisis kelompok merupakan langkah pembentukan kelompok data (kelompok) dari kumpulan data yang memiliki kelompok yang tidak diketahui sebelumnya, serta langkah menentukan penempatan data dalam kelompok yang sesuai. Proses pengelompokan bertujuan untuk mengidentifikasi kelas-kelas taksonomi atau batryologi, atau melakukan analisis topologi terhadap data yang ada. Jika dilihat dari perspektif data mining, pengelompokan tidak dapat disamakan dengan proses klarifikasi. Hal ini disebabkan karena dalam proses klarifikasi, data dikelompokkan berdasarkan kriteria pekerjaan yang telah ditentukan sebelumnya.

2.2. Tahap Clustering

Clustering merupakan proses klasifikasi menjadi beberapa bagian yang sama sesuai dengan kategori yang telah ditetapkan sebelumnya. Eunclidean Distance dapat dilakukan dengan menerapkan berbagai persamaan dan langkah-langkah mengenai jarak algoritma [5], dan untuk mendapatkan cluster sesuai dengan data yang telah dimiliki, diperlukan suatu diagram alur untuk membantu dalam alur perhitungan data yang akan diolah. Dibawah ini merupakan flowchart untuk tahapan cluster dengan K-Means [6].



Gambar 1. Diagram alur k-means

3. METODE PERANCANGAN

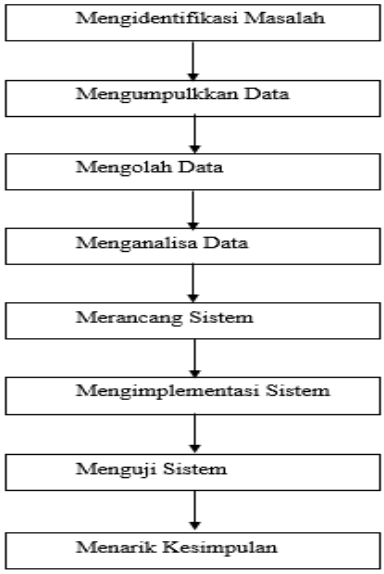
Pada bagian ini memuat metode yang digunakan pada pembuatan skripsi.

3.1. Sumber Data

Sumber data ini diperoleh dari Badan Pusat Statistik (BPS) Kota Cirebon. BPS berfungsi sebagai penyedia data untuk memenuhi kebutuhan pemerintah dan masyarakat. Data tersebut dikumpulkan melalui sensus atau survei yang dilakukan oleh BPS sendiri, serta dapat bersumber dari departemen atau lembaga pemerintahan lain sebagai data sekunder.

Data sekunder adalah data yang berhubungan dengan informasi dari sumber yang telah ada sebelumnya. Sumber data yang tidak langsung didapatkan dari objek melalui wawancara. Contoh data sekunder meliputi laporan keuangan perusahaan yang diperoleh dari data transaksi perusahaan, buku teks yang mengutip data dari sumber lain, hasil penelitian atau analisis yang diperoleh dari data primer, data sensus yang dikumpulkan oleh pemerintah, dan data pasar yang dikumpulkan oleh organisasi atau individu lain Data sekunder sering digunakan dalam penelitian karena lebih mudah diakses dan lebih murah, namun memiliki keterbatasan seperti ketidakselaluan akurasi dan kesesuaian dengan kebutuhan peneliti.

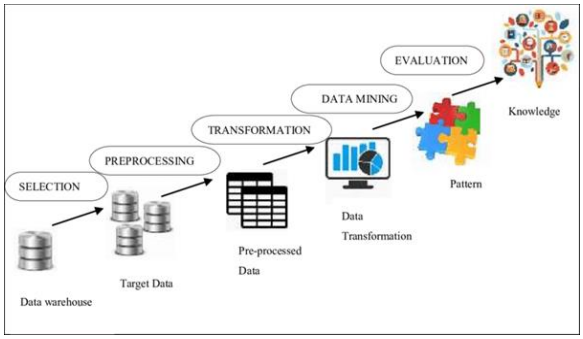
3.2. Teknik Pemrosesan Data



Gambar 2. Pemrosesan Data

3.3. Tahapan Perancangan

Tahapan-tahapan yang dilakukan dalam penelitian ini adalah menggunakan perancangan *Knowledge Discovery in Databases* (KDD). KDD dapat didefinisikan sebagai proses penggalian pengetahuan atau pola yang berguna dari kumpulan data yang besar. Tahapan KDD bisa di gambarkan seperti di bawah ini :



Gambar 3. Tahapan KDD

Berikut ini adalah langkah penting dalam menggunakan perancangan Knowledge Discovery in Databases (KDD), seperti:

a. Data *Selection*

Data *Selection* dapat didefunisikan sebagai proses sistematis untuk mengumpulkan informasi atau data yang relevan dari berbagai sumber. Proses ini mencakup identifikasi data yang dibutuhkan, menentukan metode yang tepat untuk mengumpulkannya, dan memastikan

keakuratan dan keandalan data yang terkumpul. Definisi ini mencakup berbagai teknik, termasuk survei, wawancara, observasi, eksperimen, dan ekstraksi data yang ada dari dokumen atau basis data. Tujuan dari pengumpulan data adalah untuk mendapatkan informasi yang handal dan valid yang dapat dianalisis dan diinterpretasikan untuk menjawab pertanyaan penelitian atau mencapai tujuan tertentu [7].

b. Data Preprocessing
Data *Preprocessing* adalah langkah yang sangat penting dalam proses analisis data. Proses ini melibatkan transformasi data mentah ke dalam format yang dapat dengan mudah dipahami dan dianalisis oleh model pembelajaran mesin atau algoritme penggalian data. Tujuan dari pemrosesan data adalah untuk meningkatkan kualitas data dengan mengidentifikasi dan mengoreksi kesalahan, inkonsistensi, atau nilai yang hilang. Dengan melakukan *preprocessing* data, kita dapat memastikan bahwa data tersebut dapat diandalkan, konsisten, dan siap untuk analisis lebih lanjut [8].

c. Data *Transformation*
Transformasi data didefinisikan sebagai pengacu pada proses mengubah data dari format atau struktur aslinya ke dalam format atau struktur yang berbeda. Transformasi data melibatkan manipulasi nilai data, mengatur ulang elemen data, atau memodifikasi organisasi data secara keseluruhan. Tujuan dari transformasi data adalah untuk memastikan bahwa data berada dalam kondisi atau format yang diinginkan yang memenuhi persyaratan khusus untuk analisis, integrasi, atau penyimpanan. Proses ini dapat melibatkan berbagai operasi, seperti penyaringan, penyortiran, penggabungan, penggabungan, pemisahan, dan normalisasi data. Transformasi data merupakan langkah mendasar dalam manajemen dan analisis data, yang memungkinkan organisasi memperoleh wawasan yang bermakna dan mencapai tujuan bisnis mereka [9].

d. Data Mining
Data mining didefinisikan sebagai praktik memeriksa dataset yang besar untuk mengidentifikasi pola atau hubungan yang dapat digunakan untuk membuat keputusan bisnis atau mendapatkan wawasan yang lebih dalam dari data tersebut [10]. Tujuan dari data mining adalah untuk mengungkap wawasan dan pengetahuan tersembunyi yang mungkin tidak terlihat melalui metode analisis tradisional. Hal ini melibatkan penggunaan berbagai teknik statistik dan pembelajaran mesin seperti Rapidminer atau tools lainnya yang dapat digunakan untuk mengidentifikasi pola, membuat prediksi, dan mendapatkan pemahaman yang lebih dalam tentang data. Tujuan utamanya adalah untuk mengekstrak informasi yang bermakna dan dapat ditindaklanjuti yang dapat mendorong peningkatan dan mengoptimalkan proses pengambilan keputusan [11].

e. Evaluation
Evaluation didefinisikan sebagai proses sistematis dalam mengumpulkan dan menganalisis informasi untuk menentukan nilai, manfaat, atau keefektifan suatu program, intervensi, atau proyek. *Evaluation* ini melibatkan penilaian terhadap kekuatan dan kelemahan subjek yang dievaluasi, hasil, dampak, dan kinerja secara keseluruhan. Evaluasi memberikan wawasan dan rekomendasi yang berharga untuk perbaikan, pengambilan keputusan, dan akuntabilitas. Evaluasi dapat diterapkan di berbagai bidang seperti pendidikan, kesehatan, dan kesejahteraan sosial. Proses evaluasi biasanya mencakup perencanaan, pengumpulan data, analisis, dan pelaporan, serta dipandu oleh prinsip-prinsip utama dan praktik terbaik [12].

f. *Knowledge*
Knowledge didefinisikan sebagai pemahaman, kesadaran, atau keakraban dengan fakta, informasi, keterampilan, atau konsep yang diperoleh melalui pengalaman, pendidikan, atau studi. Pengetahuan yang lebih banyak dari sekadar informasi dan melibatkan kemampuan untuk menerapkan dan menggunakan informasi tersebut secara efektif. Pengetahuan dapat bersifat individual dan kolektif, karena dapat dimiliki oleh individu atau dibagikan di antara sekelompok orang. Hal ini sering kali ditandai dengan perpaduan antara pemahaman teoritis dan aplikasi praktis, yang memungkinkan individu untuk membuat keputusan yang tepat dan memecahkan masalah di berbagai domain [8].

4. HASIL DAN PEMBAHASAN

4.1. Implementasi Penerapan Algoritma K-Means

Pada bagian hasil tugas akhir ini penulis menggunakan metode pendekatan KDD (*Knowledge Discovery in Databases*) menggunakan *tools* rapid miner. Berikut ini tahapan KDD yaitu :

4.1.1. Data *Selection*

Tahapan pertama dalam proses KDD yaitu data *selection*. Pada tahapan ini penulis melakukan seleksi data yang akan digunakan dalam proses pengelompokkan. Data yang akan diproses yaitu data penduduk se-wilayah Jawa Barat yang akan dikelompokkan berdasarkan pekerjaan. Tahapan seleksi data bertujuan untuk memilih data yang akan digunakan dalam proses pengelompokkan melalui aplikasi Rapidminer, tahapan seleksi data ini dilakukan di *Microsoft Excel*. Data yang digunakan dalam penelitian ini yaitu data pekerjaan penduduk di wilayah Jawa Barat dari tahun 2011-2023 dengan dataset total berjumlah 351 dan 7 atribut diantaranya

Id, Kode Provinsi, Nama Provinsi, Jenis pekerjaan atau jabatan, Jumlah Penduduk, Satuan, Tahun. Dataset yang digunakan dapat dilihat pada gambar 4

Row No.	ID	Kode Provinsi	Nama Provi...	Jenis Pekerj...	Jumlah Pen...	Satuan	Tahun
337	337	32	JAWA BARAT	TENAGA PEK...	1083853	ORANG	2023
338	338	32	JAWA BARAT	WIRAUSAHA	228284	ORANG	2023
339	339	32	JAWA BARAT	PEDAGANG ...	933097	ORANG	2023
340	340	32	JAWA BARAT	GURU	4093001	ORANG	2023
341	341	32	JAWA BARAT	DOKTER	1134771	ORANG	2023
342	342	32	JAWA BARAT	PEGAWAI RE...	3615440	ORANG	2023
343	343	32	JAWA BARAT	KOKI	6366335	ORANG	2023
344	344	32	JAWA BARAT	TUKANG BE...	1142616	ORANG	2023
345	345	32	JAWA BARAT	TUKANG AN...	165005	ORANG	2023
346	346	32	JAWA BARAT	SUPIR BUS	972235	ORANG	2023
347	347	32	JAWA BARAT	SUPIR TRUK	4108318	ORANG	2023
348	348	32	JAWA BARAT	PEGAWAI BA...	1183197	ORANG	2023
349	349	32	JAWA BARAT	KARYAWAN PT	3888313	ORANG	2023
350	350	32	JAWA BARAT	MANUFAKTUR	6861424	ORANG	2023
351	351	32	JAWA BARAT	INTERNET M...	1211382	ORANG	2023

ExampleSet (351 examples, 0 special attributes, 7 regular attributes)

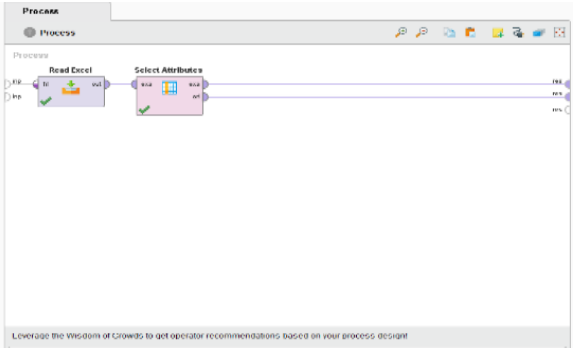
Gambar 4. Dataset Pekerjaan Penduduk

4.1.2. Data Preprocessing

Setelah melakukan tahapan data selection tahapan selanjutnya adalah melakukan data *preprocessing*. Tujuan dari data *preprocessing* yaitu untuk penghapusan atribut data yang tidak diperlukan, data yang *null* atau *missing*, menghilangkan data yang tidak konsisten, serta menambahkan id kedalam dataset yang akan digunakan. Pada tahapan data *preprocessing* ini penulis menggunakan 2 jenis operator yaitu operator *Select Atributes* dan operator *Set Role*. Adapun langkah-langkah yang dilakukan penulis dalam menggunakan 2 operator tersebut yaitu sebagai berikut :

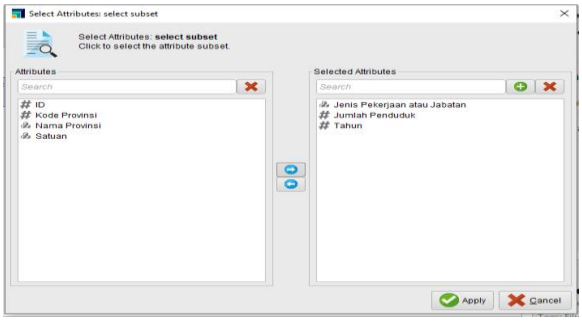
a. Operator *Select Atributes*

Langkah pertama yang dilakukan penulis dalam tahapan *processing* yaitu menggunakan operator *Select Atributes*. Operator *Select Atributes* adalah operator yang berfungsi untuk memilih *subset* atribut yang akan digunakan dari *Examplesate* dan menghapus atribut lainnya yan tidak digunakan. Tampilan operator *Select Atributes* dapat dilihat pada gambar 5



Gambar 5. Operator Select Atributes

Pada operator *Select Atributes* terdapat parameter yang harus disesuaikan terlebih dahulu. Tampilan proses mengunakan parameter *select atributes* untuk menghilangkan atribut yang tidak digunakan dapat dilihat pada gambar 6



Gambar 6. Parameter Select Atributes

Atribut data yan tidak digunakan dalam penelitian ini adalah atribut data ID, Kode Provinsi, Nama Provinsi, dan Satuan, maka setelah dilakukan proses *Select Atributes* maka dataset yang semula berjumlah 7 atribut sekarang berjumlah 3 atribut yang akan digunakan yaitu Jenis Pekerjaan atau Jabatan, Jumlah Penduduk, dan Tahun. Hasil yang diperoleh dari proses operator ini dapat dilihat pada gambar 7

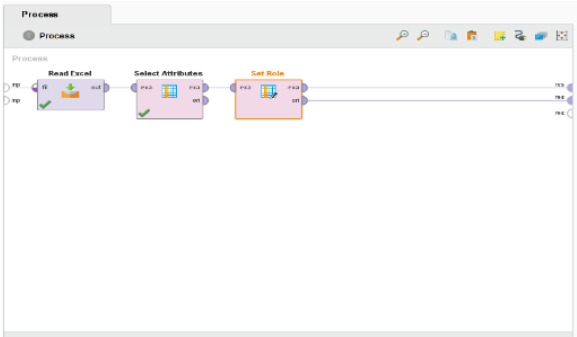
Row No.	Jenis Pekerjaan atau Jabatan	Jumlah Pen...	Tahun
1	TENAGA PROFESIONAL	1083853	2011
2	TENAGA KEPEMIMPINAN DAN KETATALAK...	228284	2011
3	TENAGA TATA USAHA	933097	2011
4	TENAGA USAHA PENJUALAN	4093001	2011
5	TENAGA USAHA JASA	1134771	2011
6	TENAGA USAHA PERTANIAN	3615440	2011
7	TENAGA PRODUKSI	6366335	2011
8	TENAGA TEKNISI	1142616	2011
9	TENAGA KEHUTANAN	165005	2011
10	TENAGA PERBURUAN	972235	2011
11	TENAGA PERIKANAN	4108318	2011
12	TENAGA OPERATOR ALAT-ALAT ANGKUT...	1183197	2011
13	TENAGA PEKERJA KASAR	3888313	2011
14	WIRAUSAHA	6861424	2011
15	PEDAGANG KAKI LIMA	1211382	2011

ExampleSet (351 examples, 0 special attributes, 3 regular attributes)

Gambar 7. Exampleset Data Setelah Proses Select Atributes

b. Operator *Set Role*

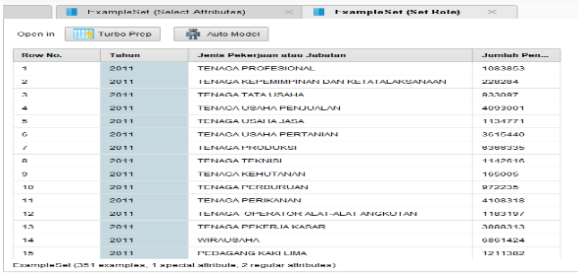
Langkah yang kedua dalam tahapan *preprocessing* adalah menggunakan operator *Set Role*. Operator *Set Role* digunakan untuk mengubah peran dari satu atau lebih atribut. Tampilan operator *Set Role* dapat dilihat pada gambar 8



Gambar 8. Operator Set Role

Pada operator *Set Role* terdapat parameter yang harus disesuaikan terlebih dahulu, dalam proses menggunakan parameter *Set Role* itu untuk mengubah atribut menjadi id. Atribut data yang

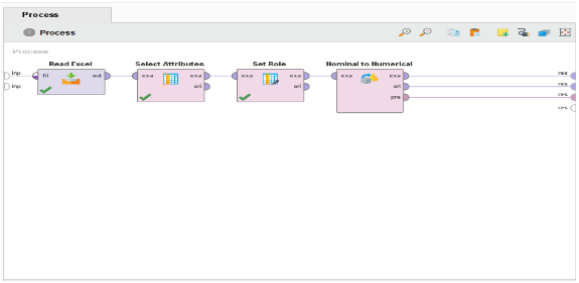
menjadi target *role* regular yaitu atribut Tahun. Hasil yang diperoleh dari proses operator ini dapat dilihat pada gambar 9



Gambar 9. Hasil Setelah Menggunakan Set Role

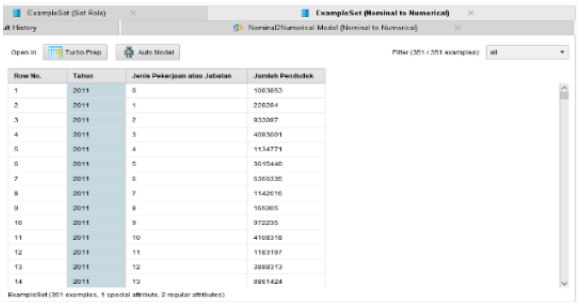
4.1.3. Data Transformation

Setelah melakukan tahapan data preprocessing maka tahapan selanjutnya adalah melakukan data transformation. Pada tahapan transformation penulis mengubah tipe atribut data yang semula *non-numerik* diubah menjadi tipe data *numerik.*Atribut data yang diubah yaitu atribut Jenis pekerjaan atau jabatan menjadi angka menggunakan operator *Nominal to Numerical* agar sesuai dengan tipe data yang dibutuhkan algoritma K-means pada prose *clustering.* Tampilan operator Nominal to Numerical dapat dilihat pada gambar 10



Gambar 10. Operator Nominal to Numerical

Dalam menggunakan operator *Nominal to Numerical* terdapat parameter yang harus disesuaikan terlebih dahulu agar sesuai dengan kebutuhan pada proses *clustering* nanti. Hasil yang diperoleh dari transformasi Jenis Pekerjaan atau Jabatan menjadi *Unique Integers* menggunakan *Nominal to Numerical* dapat dilihat pada gambar 11



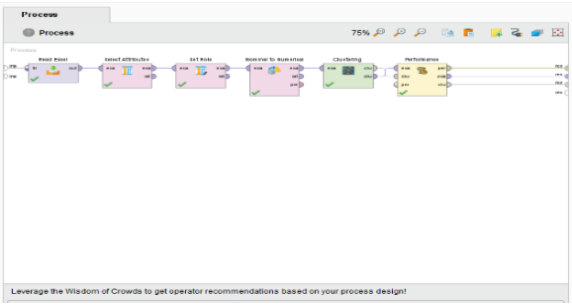
Gambar 11. Hasil Data Setelah di Transformation

4.1.4. Data Mining

Setelah melakukan tahapan data *transformation* maka tahapan selanjutnya adalah melakukan penerapan data mining. Pada tahapan penerapan data mining penulis menggunakan aplikasi Rapidminer versi 10.1.001 dengan metode algoritma k-means. Pada tahapan data mining ini penulis melakukan dua tahapan yaitu tahapan pertama penulis melakukan proses *clustering* menggunakan algoritma K-means dengan operator yang digunakan yaitu operator *Clustering* (K-Means) dan tahapan kedua penulis melakukan pengujian dan mengevaluasi hasil dari *clustering* menggunakan operator *Cluster Distance Performance* dengan metode yang digunakan evaluasi *Davies Bouldin Index* (DBI).

a. Tahapan *Clustering* K-means

Tahapan pertam yaitu penulis melakukan proses *clustering* K-means dilakukan 8 kali percobaan dimulai dari k = 3 sampai k = 10. Pemodelan data mining pada pengelompokkan data pekerjaan penduduk wilayah Jawa Barat menggunakan algoritma K-means dapat dilihat pada gambar 12



Gambar 12. Model Penerapan Algoritma K-means

b. Evaluasi *Davies Bouldin Index* (DBI)

Tahapan kedua yaitu melakukan pengujian hasil *clustering* dengan menerapkan nilai DBI pada paameter *Cluster Distance Performance.* Pada pengujian DBI, *cluster* yang memiliki nilai DBI terkecil atau mendekati 0 dijadikan sebagai kluster yang terbaik. Untuk mencari cluster terkecil penulis melakukan percobaan dari cluster 2 sampai dengan 10 rekapitulasi. Jumlah cluster yang dihasilkan dari nilai *Davies Bouldin Index* dapat dilihat dari tabel 2 berikut :

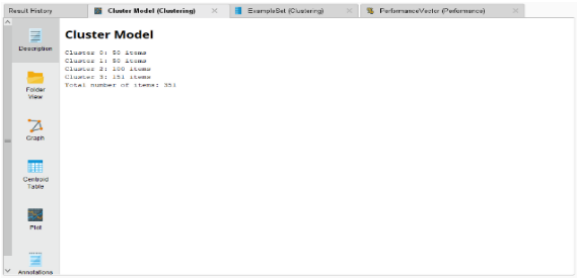
Tabel 2. Evaluasi *Davies Bouldin Index*

Clustering	Jumlah Anggota Cluster	Hasil Nilai DBI
2	Cluster 0 : 225 items	0,486
	Cluster 1 : 126 items	
3	Cluster 0 : 201 items	0,304
	Cluster 1 : 50 items	
	Cluster 2 : 100 items	
4	Cluster 0 : 50 items	0,262
	Cluster 1 : 50 items	
	Cluster 2 : 100 items	
	Cluster 3 : 151 items	
5	Cluster 0 : 50 items	0,366
	Cluster 1 : 50 items	
	Cluster 2 : 43 items	

Clustering	Jumlah Anggota Cluster	Hasil Nilai DBI
	Cluster 3 : 151 items	
	Cluster 4 : 57 items	
6	Cluster 0 : 151 items	0,369
	Cluster 1 : 38 items	
	Cluster 2 : 43 items	
	Cluster 3 : 12 items	
	Cluster 4 : 50 items	
	Cluster 5 : 57 items	
7	Cluster 0 : 38 items	0,375
	Cluster 1 : 30 items	
	Cluster 2 : 151 items	
	Cluster 3 : 50 items	
	Cluster 4 : 12 items	
	Cluster 5 : 46 items	
	Cluster 6 : 24 items	
8	Cluster 0 : 24 items	0,409
	Cluster 1 : 38 items	
	Cluster 2 : 91 items	
	Cluster 3 : 46 items	
	Cluster 4 : 12 items	
	Cluster 5 : 60 items	
	Cluster 6 : 50 items	
	Cluster 7 : 30 items	
9	Cluster 0 : 12 items	0,379
	Cluster 1 : 50 items	
	Cluster 2 : 46 items	
	Cluster 3 : 24 items	
	Cluster 4 : 31 items	
	Cluster 5 : 60 items	
	Cluster 6 : 91 items	
	Cluster 7 : 30 items	
	Cluster 8 : 7 items	
10	Cluster 0 : 24 items	0,399
	Cluster 1 : 24 items	
	Cluster 2 : 50 items	
	Cluster 3 : 38 items	
	Cluster 4 : 60 items	
	Cluster 5 : 91 items	
	Cluster 6 : 12 items	
	Cluster 7 : 14 items	
	Cluster 8 : 12 items	
	Cluster 9 : 26 items	

4.1.5. Evaluation

Pada tahapan evaluasi merupakan tahapan dalam menghasilkan pola-pola guna memperkirakan tujuan yang diharapkan. Berdasarkan tabel 4.1 menunjukkan bahwa *cluster* data pekerjaan penduduk menggunakan *Davies Bouldin Index* nilai yang paling mendekati angka 0 dengan percobaan *cluster* 2 sampai *cluster* 10 menghasilkan nilai K terbaik pada cluster 4 yaitu 0,262 dengan jumlah anggota Cluster 0 : 50 items Cluster 1 : 50 items Cluster 2 : 100 items Cluster 3 : 151 items.



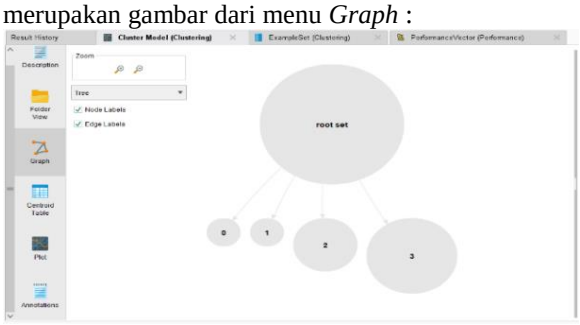
Gambar 13. Description Cluster Model K-means

4.2. Hasil 2 Analisa Hasil Penerapan Algoritma K-Means

Berikut merupakan hasil analisis data pekerjaan penduduk menggunakan penerapan algoritma K-means :

4.2.1. Hasil Analisis Dalam Bentuk Graph

Untuk memvisualisasikan data dalam berbagai bentuk menggunakan menu *Graph.* Berikut merupakan gambar dari menu *Graph* :

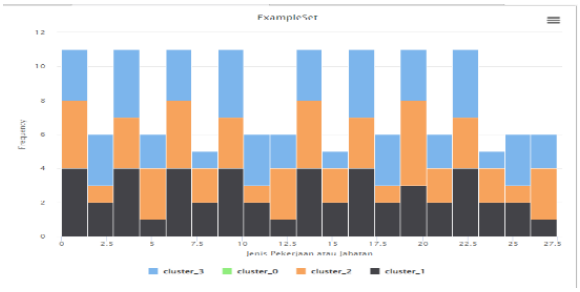


Gambar 14. Hasil Analisis Bentuk Graph

Gambar 14 merupakan ilustrasi *Graph* dalam bentuk *Tree*, semakin banyak anggota maka semakin besar bulatannya. Maka dari *Graph* tersebut dapat disimpulkan bahwa *cluster* 3 paling besar bulatannya karena memiliki jumah anggota paling banyak.

4.2.2. Hasil Analisis Dalam Bentuk *Visualization*

Berdasarkan hasil pengujian validitas *cluster* yang telah dilakukan oleh penulis maka diperoleh pengelompokkan dengan k = 4 merupakan pengelompokkan yang terbaik karena memiliki nilai DBI terkecil. Pada k = 4 jumlah anggota yang diperoleh untuk anggota Cluster 0 : 50 items, Cluster 1 : 50 items, Cluster 2 : 100 items, Cluster 3 : 151 items. Anggota dari masing-masing *cluster* dikelompokkan berdasarkan dari pekerjaan. Gambar 4.23 adalah tampilan dari proses penerapan algoritma K-means menggunakan Plot Type Histogram.



Gambar 15. Hasil Visualisasi Plot Type Histogram

5. KESIMPULAN DAN SARAN

Berdasarkan penelitian data pekerjaan penduduk di wilayah Jawa Barat dari tahun 2011-2023 dengan penggunaan 351 data dan algoritma K-means melalui Rapid Miner, kesimpulan yang dapat diambil adalah *cluster* data pekerjaan penduduk menggunakan *Davies Bouldin Index* nilai yang paling mendekati angka 0 dengan percobaan *cluster* 2 sampai *cluster* 10 menghasilkan nilai K terbaik pada cluster 4 yaitu 0,262 dengan jumlah anggota Cluster 0 : 50 items Cluster 1 : 50 items Cluster 2 : 100 items Cluster 3 : 151 items. Dan jarak antar centroid *cluster* dari data pekerjaan penduduk Jawa Barat menggunakan algoritma K-means adalah *cluster* 0 : 13.200 *cluster* 1 : 12.800 *cluster* 2 : 12.960 *cluster* 3 : 13.026.

Saran untuk penelitian mendatang adalah agar hasil klasterisasi ini dapat menjadi panduan dalam mempermudah pengelompokkan data penduduk berdasarkan pekerjaan pada tahun-tahun berikutnya. Selain itu, diharapkan penelitian ini dapat menjadi referensi yang jelas dan mendukung penelitian lebih lanjut yang lebih rinci dan terperinci.

DAFTAR PUSTAKA

[1] I. Suputra, I. Candiasa, dan I. Suryawan, "Klasterisasi Hasil Ujian Nasional SMA/MA dengan Algoritma K-Means," *Wahana Mat. dan Sains J. Mat. Sains, dan Pembelajarannya*, vol. 15, no. 1, hal. 22–30, 2021, [Daring]. Tersedia pada: https://ejournal.undiksha.ac.id/index.php/JPM/article/view/25380

[2] W. Aprianti dan J. Permadi, "K-Means Clustering untuk Data Kecelakaan Lalu Lintas Jalan Raya di Kecamatan Pelaihari," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, hal. 613–620, 2018, doi: 10.25126/jtiik.2018551113.

[3] S. X. Sui, L. J. Williams, K. L. Holloway-Kew, N. K. Hyde, dan J. A. Pasco, "Skeletal muscle health and cognitive function: A narrative review," *Int. J. Mol. Sci.*, vol. 22, no. 1, hal. 1–21, 2021, doi: 10.3390/ijms22010255.

[4] M. Ahmed, R. Seraj, dan S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electron.*, vol. 9, no. 8, hal. 1–12, 2020, doi: 10.3390/electronics9081295.

[5] S. Index, "Penerapan Algoritma K-Means Untuk Clustering Penilaian Dosen Berdasarkan Indeks Kepuasan Mahasiswa," vol. 16, no. 1, hal. 17–24, 2017.

[6] M. G. Sadewo, A. P. Windarto, dan D. Hartama, "PENERAPAN DATAMINING PADA POPULASI DAGING AYAM RAS PEDAGING DI INDONESIA BERDASARKAN PROVINSI MENGGUNAKAN K-MEANS," hal. 60–67, 2016.

[7] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, dan R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, hal. 1283–1291, 2022, doi: 10.1007/s11135-021-01176-w.

[8] S. Sieranoja dan P. Fränti, "Adapting k-means for graph clustering," *Knowl. Inf. Syst.*, vol. 64, no. 1, hal. 115–142, 2022, doi: 10.1007/s10115-021-01623-y.

[9] P. M. Weilbacher *et al.*, "The data processing pipeline for the MUSE instrument," *Astron. Astrophys.*, vol. 641, hal. 1–30, 2020, doi: 10.1051/0004-6361/202037855.

[10] V. Plotnikova, M. Dumas, dan F. Milani, "Adaptations of data mining methodologies: A systematic literature review," *PeerJ Comput. Sci.*, vol. 6, hal. 1–43, 2020, doi: 10.7717/PEERJ-CS.267.

[11] H. Priyatman, F. Sajid, dan D. Haldivany, "Klasterisasi Menggunakan Algoritma K-Means Clustering untuk Memprediksi Waktu Kelulusan Mahasiswa," *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 1, hal. 62, 2019, doi: 10.26418/jp.v5i1.29611.

[12] Solichin, Achmad, Khairunnisa, dan Khansa, "Klasterisasi Persebaran Virus Corona (Covid-19) Di DKI Jakarta," *Fountain Informatics J.*, vol. 5, no. 2, hal. 52–59, 2020.