Seminar Nasional Sains dan Teknologi Informasi (SENSASI)
SENSASI 2021
ISBN: 978-623-93614-6-4
Agustus 2021
Hal: 143 - 147

Clustering Pengeluaran Tahunan Berbagai Macam Produk Menggunakan Metode K-Means

Nisha Ananda, Rezty Amalia Aras

Bisnis Digital, Institut Teknologi dan Bisnis Kalla, Makassar, Indonesia
Email: [1]nishaananda@kallabs.ac.id, [2]reztyamalia@kallabs.ac.id

Abstrak

Data mining adalah suatu pengetahuan yang digunakan untuk mencari informasi dalam sebuah data. Data mining memiliki banyak metode untuk mengolah sebuah data dimana salah satu metode yang dapat digunakan untuk pengolahan datanya adalah dengan metode Clustering. Tulisan ini membahas tentang pemodelan data set Wholesale Customers dimana dataset tersebut berisi data distributor yang mencakup pengeluaran tahunan terhadap berbagai macam produk. Untuk pemodelan dataset dilakukan dengan metode clustering dengan menggunakan K-Means. Tujuan pengolahan data ini adalah untuk menentukan pengelompokkan yang paling tepat untuk data Wholesale Customers.

Kata Kunci: Data mining, Clustering, K-Means, Informasi, Teknologi

1. PENDAHULUAN

Pesatnya perkembangan teknologi komputer dan teknologi komunikasi membawa masyarakat ke dalam era informasi dimana dengan kematangan database dan popularitas aplikasi data. Kita hidup di dunia dimana jumlah data yang dikumpulkan oleh manusia semakin cepat dan semakin banyak setiap harinya dan menganalisis data tersebut merupakan sebuah kebutuhan yang sangat penting karena dalam menghadapi jumlah data yang semakin besar dengan beberapa struktur data, tidak semua data tersebut memiliki pengetahuan yang dibutuhkan. Oleh karena itu dibutuhkan sebuah teknologi yang dapat membantu menggali dan menganalisis data yang sangat besar tersebut untuk mendapatkan informasi yang dibutuhkan yang nantinya akan berguna dalam proses pengambilan keputusan. Memahami data yang akan diolah dalam data mining sangat penting karena pemahaman tersebut bisa menjadi kunci berhasil atau tidaknya proses data mining yang akan dilakukan.

Dengan memahami data yang akan diolah, kita bisa menentukan mulai dari input yang diperlukan untuk pengolahan data, output apa yang bisa kita peroleh dari proses data mining, perlu atau tidaknya melakukan pre-processing terhadap data yang kita miliki, apakah seluruh data telah lengkap ataukah ada nilai-nilai yang hilang (missing value), apakah setiap instance sudah memiliki atribut yang sama, dan lain sebagainya [1].Dalam tulisan ini akan dibahas sebuah proses clustering dengan metode K-Means untuk mengetahui cluster yang tepat untuk pengelompokkan pengeluaran tahunan pada dataset Wholesale Customers

2. METODOLOGI PENELITIAN

2.1 Dataset

Dataset yang digunakan bersumber dari kumpulan data pelanggan grosir yang diperoleh dari UCI Machine Learning Repository. Data ini mencakup pengeluaran tahunan terhadap berbagai macam produk. Ada 8 atribute dan 440 instances yang dimiliki oleh dataset ini. Parameter dari attribute tersebut adalah :

a. FRESH (pengeluaran tahunan pada produk segar) dengan tipe data numerik
b. MILK (pengeluaran tahunan pada produk susu) dengan tipe data numeric
c. GROCERY (pengeluaran tahunan pada produk bahan makanan) dengan tipe data numeric
d. FROZEN (pengeluaran tahunan pada produk beku) dengan tipe data numeric
e. DETERGENTS_PAPER (pengeluaran tahunan pada produk sabun dan kertas) dengan tipe data numeric
f. DELICATESSEN (pengeluaran tahunan pada produk toko makanan) dengan tipe data numeric
g. CHANNEL (pelanggan dari industri dan eceran) dengan tipe data numeric
h. REGION (pelanggan dari daerah) dengan tipe data numeric

Dari tersebut memiliki total 8 attribute dengan 6 merupakan pengeluaran tahunan sedangkan 2 attribute lainnya menunjukkan darimana customer berasal. Untuk atribut channel terbagi menjadi 2 kelas yaitu industri dan eceran. Sedangkan untuk atribut region dibagi menjadi 3 kelas yaitu Lisnon, Oporto dan Other. Penggunaan clustering sangat luas. Dalam proses komersial, pengelompokan memungkinkan untuk menganalisis pasar dengan membedakan kelompok konsumen dan untuk mensintesis pola atau kebiasaan konsumsi masing-masing jenis konsumen. Dalam data mining, dapat digunakan sebagai alat untuk menemukan beberapa informasi mendalam dalam database dan generalisasi karakteristik masing-masing kategori ataupun fokus pada kelas khusus untuk analisis lebih lanjut. Selain itu, analisis clustering dapat digunakan sebagai langkah preprocessing algoritma analisis data mining lainnya [2].

2.2 Metode Clustering K-Means

K-Means merupakan salah satu metode dalam fungsi clustering atau pengelompokan. Menurut Larose, clustering mengacu pada pengelompokan data, pengamatan atau kasus sesuai dengan kesamaan subjek yang diteliti [3].

ISBN: 978-623-93614-6-4
Agustus 2021
Hal: 143 - 147

Pengelompokkan data sesuai dengan tingkat kemiripan diantara data-data tersebut yang kemudian akan dibagi menjadi beberapa kelompok sehingga objek yang mirip akan menjadi sebuah set ini disebut dengan clustering [4]. Sehingga clustering merupakan kumpulan dari satu set objek data yang mirip antara satu dengan yang lainnya namun sama sekali berbeda antara satu kelompok dengan kelompok yang lainnya. Analisis clustering biasanya digunakan untuk mencari link berharga antara objek data dari satu set data yang diberikan. Dalam banyak aplikasi, semua objek dalam cluster sering diperlakukan sebagai objek yang akan diproses atau dianalisis

Tahapan dalam melakukan data mining salah satunya adalah preprosesing data. Artinya, data harus dibersihkan sebelum diolah. Hal ini biasanya terjadi karena data yang akan digunakan tidak cocok. Teknik atau metode yang digunakan dalam data preprocessing, diantaranya:

a. Data Cleaning
   Data cleaning adalah menghapus nilai-nilai data yang salah, mengoreksi kekacauan data dan memeriksa jika ada data yang tidak konsisten. Ada beberapa teknik untuk membersihkan data, seperti melengkapi missing value dan mengidentifikasi atau menghapus outlier.
b. Missing value
   Missing Value adalah informasi yang tidak tersedia untuk sebuah objek (kasus). Missing value terjadi karena informasi tentang suatu objek tidak tersedia, sulit ditemukan, atau informasi tersebut memang tidak ada. Pada dasarnya missing value tidak bermasalah untuk keseluruhan data, apalagi jika hanya dalam jumlah kecil, misalnya hanya 1 % dari data secara keseluruhan. Namun jika data yang hilang cukup besar, maka perlu dilakukan pengujian apakah data yang terdapat banyak missing value tersebut masih bisa diproses lebih lanjut atau tidak.
c. Data integrasi dan Data Transformasi
   Menggabungkan data dari beberapa sumber (database, data cube, atau file) ke dalam penyimpanan data yang sesuai. Data transformasi adalah tahap normalisasi dan pengumpulan data sehingga menjadi sama.

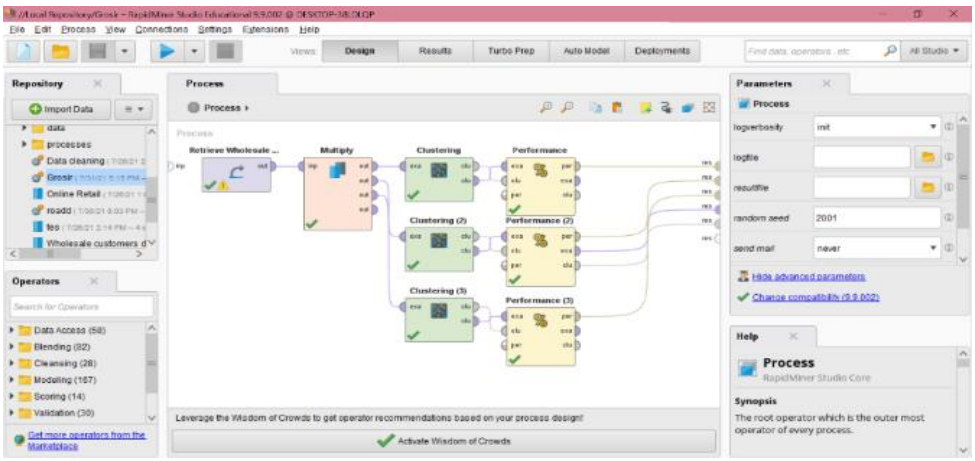
3. HASIL DAN PEMBAHASAN

RapidMiner merupakan software/perangkat lunak untuk pengolahan data. RapidMiner menggunakan prinsip dan algoritma data mining untuk mengekstrak model dari kumpulan data besar menggunakan kombinasi metode statistik, kecerdasan buatan, dan basis data. RapidMiner memungkinkan pengguna menghitung data dalam jumlah besar dengan mudah menggunakan operator. Operator ini digunakan untuk mengubah data. Data ditautkan ke masing-masing operator, sehingga hasilnya dapat dilihat hanya dengan menautkan ke operator hasil. Hasil yang ditampilkan oleh RapidMiner juga dapat ditampilkan secara grafis. Hal ini membuat RapidMiner menjadi salah satu perangkat lunak terbaik untuk mengekstraksi data dengan menggunakan teknik data mining.
Missing value tidak terdapat dalam dataset ini.

Tabel 1. Statistic Atribut pada Dataset

Atribut	Nilai Min	Nilai Maks	Nilai Rerata	Standar Deviasi
Fresh	3	112151	12000.3	12647.3
Milk	55	73498	5796.27	7380.4
Grocery	3	92780	7951.28	9503.2
Frozen	25	60869	3071.93	4854.7
Detergent	3	40827	2881.49	4767.8
Delicatess	3	47943	1524.87	2820.10

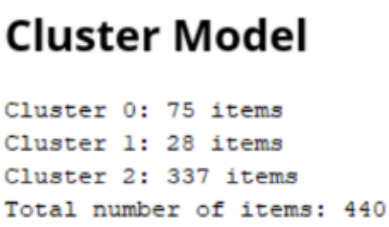
Dilihat dari nilai statistik (nilai minimum, nilai maksimum, rerata dan standar deviasi) masing-masing attribute, maka disimpulkan tidak adanya noise ataupun outlier dari dataset tersebut sehingga semua instances dapat digunakan untuk clustering. Adapun konfigurasi operator k-means pada aplikasi rapidminer sebagai berikut;



Gambar 1. Konfigurasi Operator K-Means

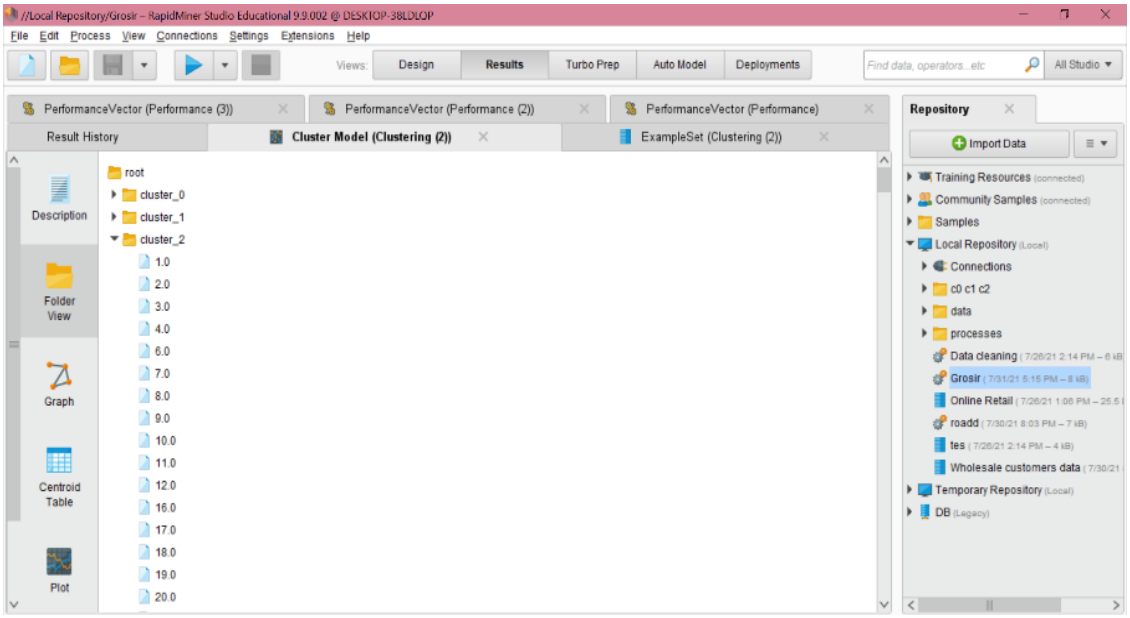
ISBN: 978-623-93614-6-4
Agustus 2021
Hal: 143 - 147

Pada Cluster Model (Clustering) dapat dilihat beberapa tampilan hasil cluster, yaitu Text View yang merupakan tampilan hasil pengelompokan berdasarkan cluster dan jumlah anggotanya.



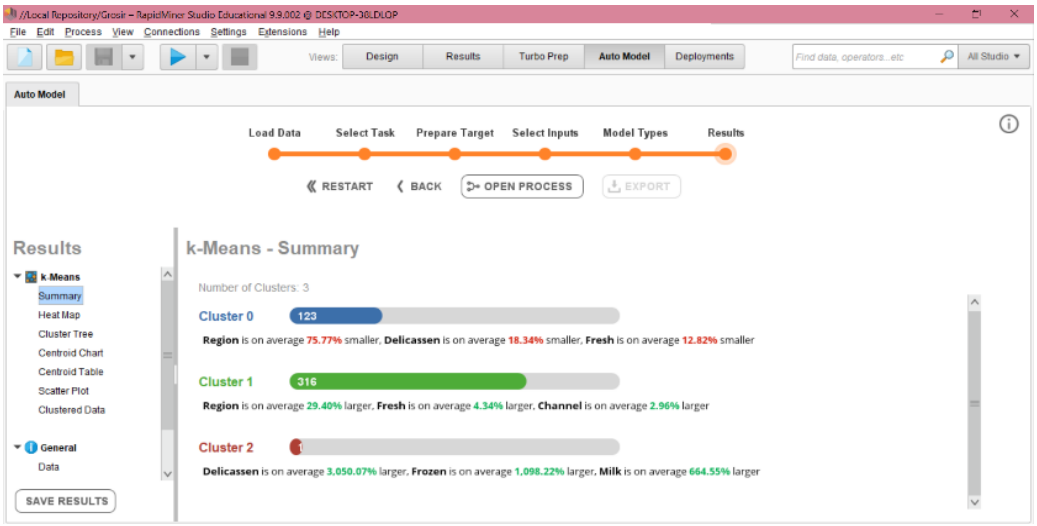
Gambar 2. Cluster Model

Pada gambar dapat kita lihat data berupa : cluster 0 yang memiliki kategori pengeluaran tahunan cukup besar memiliki 75 data. Cluster 1 yang memiliki kategori pengeluaran tahunan terbesar memiliki 28 data. Cluster 2 yang memiliki kategori terkecil memiliki 337 data.



Gambar 3. Tampilan Folder View

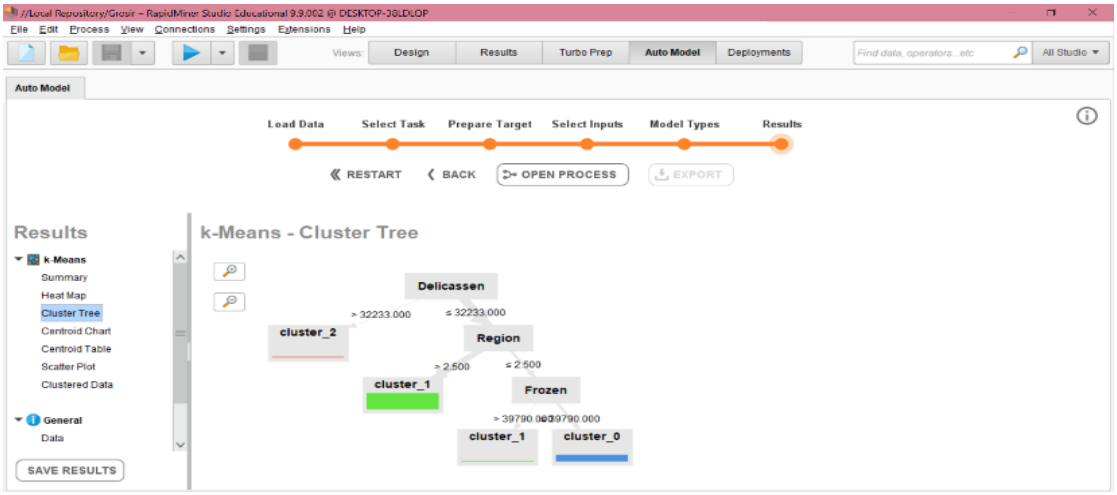
Folder View merupakan tampilan data bagian - bagian cluster secara keseluruhan, dimana masing – masing anggota cluster menampilkan field.



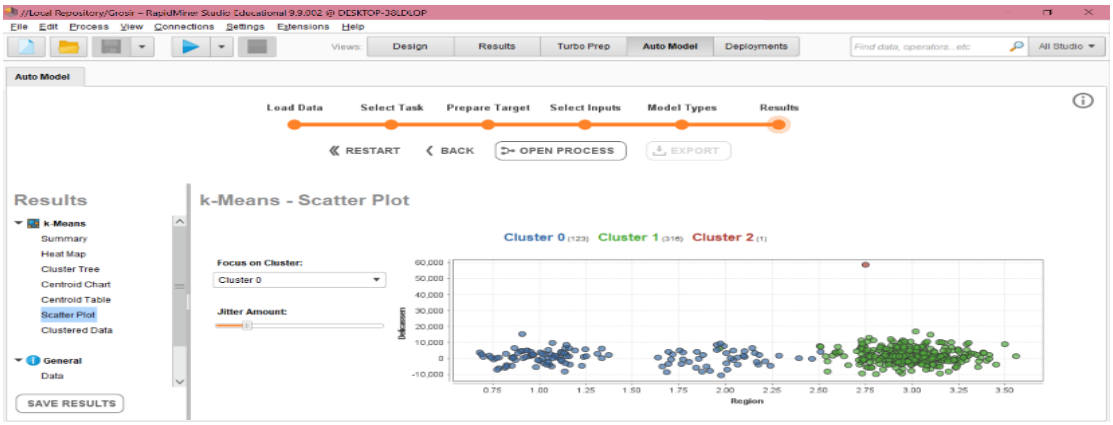
Gambar 4. K-Means Summary

ISBN: 978-623-93614-6-4
Agustus 2021
Hal: 143 - 147

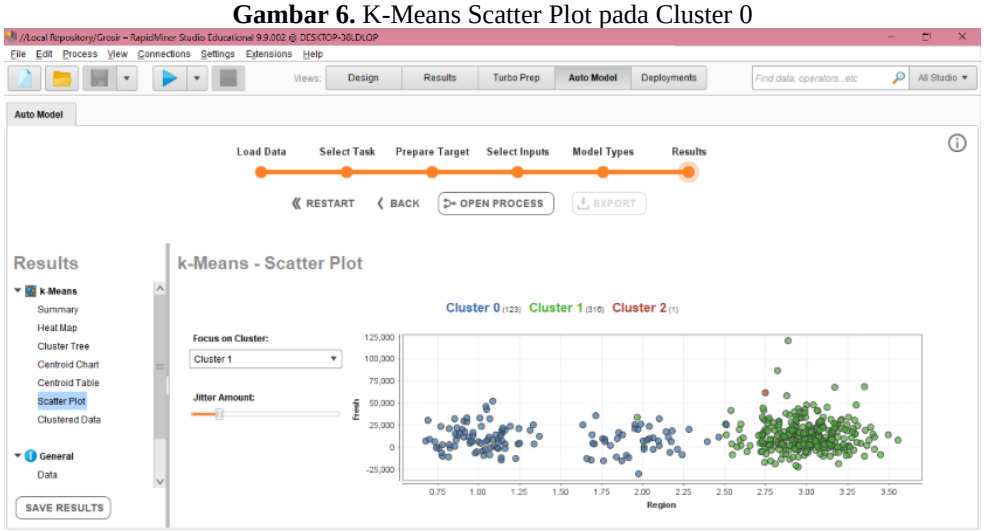
Cluster 1 sebagai cluster dengan kategori pengeluaran tahunan terbesar memiliki rata-rata Region 29.40% lebih besar, rata-rata Fresh 4.34% lebih besar, dan rata-rata Channel 2.96% lebih besar
Cluster 0 sebagai cluster dengan kategori pengeluaran tahunan cukup besar memiliki rata-rata Region 75.77% lebih kecil, rata-rata Delicassen 18.34% lebih kecil, dan rata-rata Fresh 12.82% lebih kecil
Cluster 2 sebagai cluster dengan kategori pengeluaran tahunan cukup besar memiliki rata-rata Delicassen



Gambar 5. K-Means Cluster Tree



Gambar 6. K-Means Scatter Plot pada Cluster 0



Gambar 7. K-Means Scatter Plot pada Cluster 1

4. KESIMPULAN

Berdasarkan uraian yang sudah dikemukakan sebelumnya, maka dapat ditarik kesimpulan Metode Clustering dengan menggunakan algoritma K-Means dapat digunakan untuk mengelompokkan pengeluaran tahunan berdasarkan besarnya,

ISBN: 978-623-93614-6-4
Agustus 2021
Hal: 143 - 147

yaitu terbesar, sedang, dan terkecil. Sehingga pihak toko grosir dapat mengetahui ragam produk apa yang memiliki pengeluaran tahunan terbesar. Peneliti menyadari adanya kekurangan dalam penulisan ini, karena keterbatasan penulis baik dalam hal waktu maupun pengetahuan. Dalam rangka memperbaiki kekurangan dan untuk penyempurnaan penelitian ini penulis memberikan beberapa saran.

REFERENCES

[1] Alfian, T., Sandi, A., Raharjo, M., & Putra, J. L. (2018). Clustering Kesetiaan Pelanggan E-Retail dengan Model RFM, 14, 239–246. Retrieved from https://ejournal.nusamandiri.ac.id/index.ph p/pilar/article/view/74

[2] M.Hasyim Siregar, S.Kom., M. K. (2018). Klasterisasi Penjualan Alat-alat Bangunan Menggunakan Metode K-Means, 1, 83–91. Retrieved from https://ejournal.uniks.ac.id/index.php/JTO S/article/view/24

[3] Mardalius. (2018). Pengelompokkan Data Penjualan Aksesoris Menggunakan Algoritma K-Means, IV(2), 401–411. Retrieved from https://www.researchgate.net/publication/330609314

[4] Normah, N., Rifai, B., & Sari, P. (2020). Algoritma Apriori Sebagai Solusi Kontrol Persediaan Suku Cadang Mobil PT. Buanasakti Aneka Motor Jakarta. Paradigma - Jurnal Komputer Dan Informatika, 22(2), 161–168. https://doi.org/10.31294/p.v22i2.6530

[5] Siregar, M. H. (2018). Data Mining Klasterisasi Penjualan Alat-Alat Bangunan Menggunakan Metode K-Means (Studi Kasus Di Toko Adi Bangunan). Jurnal Teknologi Dan Open Source, 1(2), 83–91. https://doi.org/10.36378/jtos.v1i2.24

[6] Yulianti, Y., Utami, D. Y., Hikmah, N., & Hasan, F. N. (2019). Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Mengetahui Minat Customer Di Toko Hijab. Jurnal Pilar Nusa Mandiri, 15(2), 241–246. https://doi.org/10.33480/pilar.v15i2.65