Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2022
Institut Informatika dan Bisnis Darmajaya, 27 Agustus 2022
ISSN: 2598-0256, E-ISSN: 2598-0238

Implementasi K-Means Dalam Mengatasi Masalah *Cold Star* Pada *Collaborative Filtering*

Sylvia Sylvia [1a], Sri Lestari[2b]

[ab] Magister Teknik Informatika, Institut Informatika dan Bisnis Darmajaya Lampung
[a] *sylviamkmr@gmail.com*
[b] *srilestari@darmajaya.ac.id*

Abstract

The development of the recommendation system is currently growing rapidly. Recommendation system is a technology that can recommend a certain item to users. One method that is often used is collaborative filtering. The workings of collaborative filtering is to provide recommendations to users based on the assessments of other users. However, there is a problem found, namely the presence of new users (cold stars) that affect the performance of the recommendation system, so that the recommendation system has difficulty analyzing the direction of user interest where new users have not rated a product which results in the system not being able to recommend something. Therefore, an algorithm is needed to alleviate the cold star problem with a clustering approach using K-Means. Research on the clustering process uses data from Movielens.com, through the selection of the k-means attribute, dividing the demographic data into 15 clusters.

Keywords*: Clustering; K-Means; Rapid Miner; Cold Star*

Abstrak

Perkembangan sistem rekomendasi saat ini berkembang secara pesat. Sistem rekomendasi merupakan teknologi yang dapat merekomendasikan suatu item tertentu kepada pengguna. Salah satu metode yang kerap digunakan ialah *collaborative filtering*. Cara kerja dari *collaborative filtering* adalah dengan memberikan rekomendasi kepada user berdasarkan penilaian dari pengguna lain. Namun, ada sebuah masalah yang ditemukan yaitu adanya pengguna baru (*cold star*) yang mempengaruhi kinerja dari sistem rekomendasi, sehingga sistem rokomendasi mengalami kesulitan untuk menganalisa arah peminatan pengguna dimana pengguna baru belum memberikan rating terhadap suatu produk yang mengakibatkan tidak dapatnya suatu sistem merekomendasikan sesuatu. Maka dari itu dibutuhkan suatu algoritma untuk meringankan masalah *cold star* dengan pendekatan *clustering* menggunakan K-Means. Penelitian proses clustering menggunakan data dari Movielens.com, melalui seleksi atribut k-means membagi data demografi menjadi 15 *cluster*.

Keywords: Klastering; K-Means; Rapid Miner; Cold Star

1. PENDAHULUAN

Pemanfaatan teknologi informasi telah merambah di berbagai bidang kesehatan, perbankan, pendidikan dan juga dunia perfilman. Teknologi informasi dapat berkembang dengan pesat karena berbagai kemudahan yang diberikan, seperti kemudahan mendapatkan dan menyebarkan informasi, serta peningkatan layanan personal. Peningkatan layanan dilakukan dengan memanfaatkan berbagai teknik dan media agar perusahaan mampu bersaing, salah satunya dengan sistem *on-line*. Kelebihan dari sistem *on-line* yaitu mampu memasarkan produk dengan cepat, tanpa terkendala tempat dan waktu. Namun seiring berjalannya waktu dan berkembangnya perusahaan maka semakin banyak dan beragam jumlah produk yang dipasarkan. Hal tersebut berdampak kepada pengguna yang mengalami kebingungan dalam memilih produk yang diinginkan sehingga menyita waktu(Hu, 2014). Film merupakan suatu media hiburan yang popular di kalangan masyarakat Indonesia. Menurut laporan kementrian Pendidikan dan Kebudayaan ada lebih dari 3423 film yang telah di produksi Indonesia di tahun 2022. Saat ini dapat dikatakan bahwa antusias masyarakat untuk menonton film tengah menggeliat bangun. Dapat di tandai dengan begitu massive nya jumlah produksi film dari tahun ketahun. Disamping itu, film Hollywood tidak kalah menggencar perfilman indonesia. Dengan begitu banyak nya variasi film dan genre yang begitu luas tentu membuat masyarakat mengalami kebingungan untuk memilih rekomendasi film apa saja yang harusnya di tonton. Ketersediaan data mengenai peringkat film dapat dimanfaatkan dan digunakan untuk merekomendasikan film kepada pengguna lain. Keterkaitan antar film yang telah di rating dapat dijadikan suatu rekomendasi kepada pengguna lain. Oleh karena itu, diperlukan suatu sistem yang dapat merekomendasikan film kepada pengguna.

Untuk mengatasi masalah dalam melakukan rekomendasi, salah satu solusi yang digunakan adalah dengan membangun sistem rekomendasi dengan menggunakan metode *Collaborative filtering*. *Collaborative filtering* memiliki beberapa kelebihan diantaranya mudah diimplementasikan dan dapat menyaring segala jenis informasi atau

ISSN: 2598-0256, E-ISSN: 2598-0238

barang tanpa harus menganalisis komentar-komentar dari pengguna. Selain itu *collaborative filtering* menghasilkan rekomendasi kualitas tinggi daripada sistem rekomendasi berdasarkan konten dan demografis(Lien & Phuong, 2014). *Collaborative filtering* mempelajari perilaku, aktifitas dan juga kencenderungan pengguna. *Cold start* adalah suatu kondisi di mana pengguna baru yang belum pernah memberikan *rating* terhadap suatu produk, sehingga informasi yang didapatkan untuk arah peminatan dari pengguna sulit diketahui(Huang & Dai, 2015). Dengan terlalu sedikitnya informasi dari pengguna baru, sistem rekomendasi mengalami kesulitan untuk memprediksi dan merekomendasikan suatu film kepada pengguna tersebut. Oleh karena itu maka penelitian ini mengusulkan pendekatan *clustering* untuk meringankan permasalahan yang di alami oleh pengguna baru dengan memanfaatkan algoritma clustering K-Means.

2. KERANGKA TEORI

2.1. Collaborative Filtering

Collaborative filtering merupakan salah satu metode rekomendasi yang menggunakan data rating dari pengguna lain untuk menghasilkan rekomendasi. Metode tersebut menganggap bahwa selera pengguna terhadap suatu produk akan cenderung sama dari waktu ke waktu, begitu pula dengan pengguna lain yang memiliki selera sama. Penggunaan collaborative filtering terbukti sangat baik(Laksana, 2014). Pada algoritma ini, memungkinkan analisa atau data preferensi untuk di kelompokan sesuai dengan kesamaan selera menjadi beberapa kelompok sehingga dapat membantu mengambil keputusan. *Collaborative filtering* memiliki kelebihan dimana dapat mengatasi dan bekerja dengan baik dalam konten yang sulit dianalisis sekalipun.

2.2. Cold Star

Cold start adalah suatu kondisi pengguna baru yang belum pernah memberikan rating terhadap suatu produk, sehingga informasi yang didapatkan untuk arah peminatan pengguna sulit diketahui(Ruswanda et al., 2015). Masalah *sparity* terjadi ketika pengguna baru memasukin sistem, sebab degan informasi yang sangat terbatas tentang pengguna tersebut. Biasanya, hasil evaluasi tidak berguna dan condong pada rekomendasi yang lemah (Al-Bashiri et al., 2017). Sebagai contoh, pengguna baru tidak dapat merekomendasikan film karena tidak adanya pemgguna dalam sistem yang serupa dengan pengguna baru tersebut.

2.3. K-Means Clustering

Clustering merupakan salah satu metode yang digunakan untuk menganalisa data, dengan tujuan untuk mengelompakkan data sesuai dengan karakteritik yang sama. K-means merupakan algoritme cluster yang banyak digunakan oleh peneliti, karena merupakan algoritme yang sederhana dan prosesnya cepat(Min & Shuzhen, 2015). Salah satu metode yang digunakan untuk menghitung jarak Euclidean Distance Space dengan mengetahui jarak terpendek antara dua titik. Perhitungan jarak menggunakan Persamaan (1)

$$D(x_i, c_j) = \sqrt{\sum_{i=1}^{N} (x_i - c_j)^2} \quad (1)$$

Dimana :
D (i, j) = Jarak data ke i ke pusat cluster j
xi = Data ke i pada atribut data ke k
cj = Titik pusat ke j pada atribut ke k

2.4. Rapid Miner

Penelitian ini menggunakan applikasi Rapid Miner. Rapid miner merupakan sebuah perangkat lunak yang dikembangkan oleh perusahaan dengan nama yang sama yang menyediakan integrase lingkungan untuk pembelajaran mesin, penambangan data, penambangan teks analitik prediktif dan juga bisnis(Tripathi et al., 2016). Aplikasi ini dipilih sebab Rapid Miner dapat menyertakan visualisasi, statistik, dan informasi terpenting tentang model ke dalam laporan dengan sekali klik. Rapid Miner menyertakan pelaporan cerdas di mana pengguna dapat mengakses riwayat alur kerja untuk setiap widget dan visualisasi langsung dari laporan. Bukan hanya itu, applikasi ini memiliki antar muka yang baik, sehingga pengguna hanya focus pada analisis data bukan pada pengkodean. Membuat konstruksi pipeline analisis data yang kompleks menjadi sederhana.

Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2022
Institut Informatika dan Bisnis Darmajaya, 27 Agustus 2022
ISSN: 2598-0256, E-ISSN: 2598-0238

3. METODOLOGI

3.1. Data Set

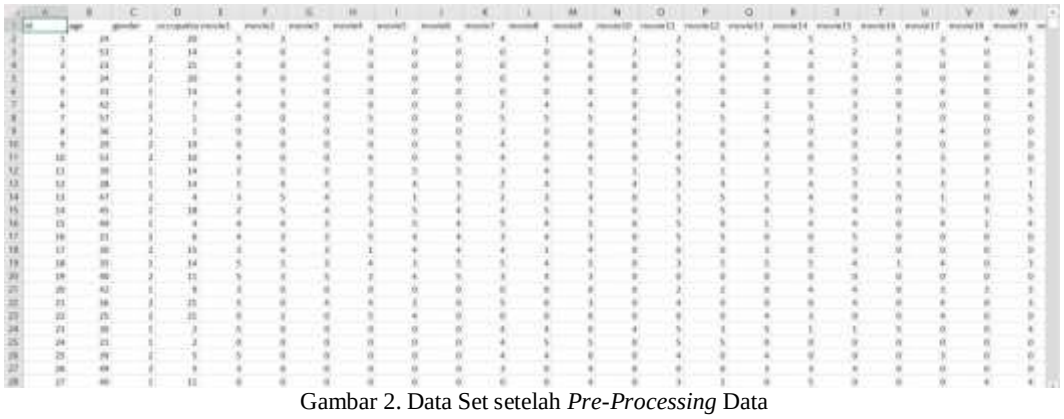
Pengumpulan data dalam penelitian ini di ambil melalui web Movilens.umn.edu. Kumpulan data terdiri dari 100.000 peringkat dengan skala 1-5 dari 943 pengguna di 1682 film. Setiap pengguna setidaknya telah menilai 20 film dengan informasi demografis pengguna adalah usia, jenis kelamin, pekerjaan dan juga kode pos. Baik Universitas Minnesota maupun peneliti yang telah terlibat menjamin kebeneran data dan keabsahannya.

3.2. Tahap Pengolahan Data

Data yang telah diperoleh akan melalui proses pre-procesing data, dimana data yang ada mengalami tahap *cleansing data* dan transformasi data. Pada tahap *cleansing data*, *record data* yang tidak konsisten akan dihapus untuk mendapatkan hasil data berkualitas tinggi(Darwis et al., 2021) sedangkan data yang memiliki rentang nilai berbeda akan di normalisasi untuk menyamakan atribut dengan skala tertentu agar mendapatkan hasil data mining yang baik(Nasution et al., 2019) sehingga kolom *age*, *gender* dan *occopution* dapat diganti dengan nilai. Sehingga pada tahap ini data telah dapat diproses di tahap *clustering* menggunakan algoritma K-Means(Wulan Sari et al., 2018).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	age	gender	occupation	movie1	movie2	movie3	movie4	movie5	movie6	movie7	movie8	movie9	movie10
2	1	24	M	technician	5	3	4	3	3	5	4	1	5	3
3	2	53	F	other	4	0	0	0	0	0	0	0	0	2
4	3	23	M	writer	0	0	0	0	0	0	0	0	0	0
5	4	24	M	technician	0	0	0	0	0	0	0	0	0	0
6	5	33	F	other	4	3	0	0	0	0	0	0	0	0
7	6	42	M	executive	4	0	0	0	0	0	2	4	4	0
8	7	57	M	administrato	0	0	0	5	0	0	5	5	5	4
9	8	36	M	administrato	0	0	0	0	0	0	3	0	0	0
10	9	29	M	student	0	0	0	0	0	5	4	0	0	0
11	10	53	M	lawyer	4	0	0	4	0	0	4	0	4	0
12	11	39	F	other	2	5	5	5	5	5	3	4	5	1
13	12	28	F	other	1	4	4	3	4	3	2	4	1	4
14	13	47	M	educator	3	5	4	2	1	2	2	3	4	0
15	14	45	M	scientist	2	5	4	5	5	4	4	5	3	0
16	15	49	F	educator	4	4	3	3	5	4	5	4	5	0
17	16	21	M	entertainme	4	3	2	5	4	4	3	4	3	0

Gambar 1. Data Set sebelum *Pre-Processing* Data

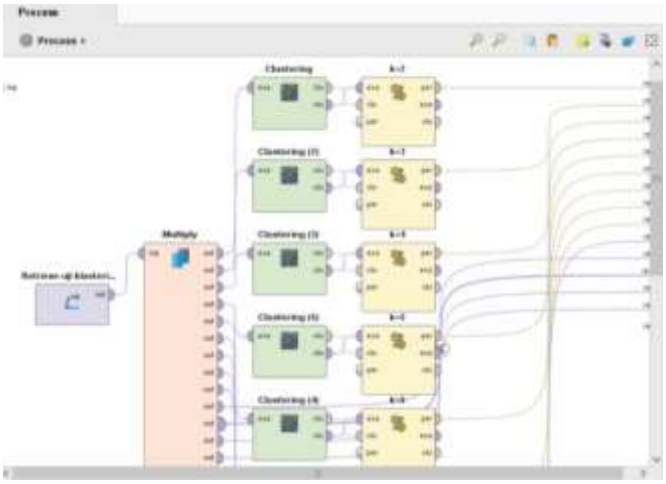


Gambar 2. Data Set setelah *Pre-Processing* Data

3.3. Tahap Clustering

Data yang telah diperoleh akan melalui proses pre-procesing data. Dimana data yang ada mengalami tahap *cleansing data* dan transformasi data, sehingga pada tahap ini telah diperoleh perhitungan yang akan diproses di tahap *clustering* menggunakan algoritma K-Means(Wulan Sari et al., 2018). Untuk mendapatkan nilai K optimal, penulis menggunakan Davies Bouldine Index yang merupakan sebuah metode untuk memvalidasi cluster evaluasi kuantitatif dari hasil clustering dengan berlandaskan nilai kohesi dan separasi(E. Prasetyo, 2014). Adapun nilai yang dipilih adalah nilai paling kecil yang diperoleh oleh DBI (non negative >=0), maka semakin kecil nilai dari uji DBI semakin baik performa cluster K-Means.

ISSN: 2598-0256, E-ISSN: 2598-0238



Gambar 3. Proses Clustering

Setelah dilakukan percobaan, ditemukan nilai DBI seperti yang ditanpilkan pada tabel 1. dibawah ini

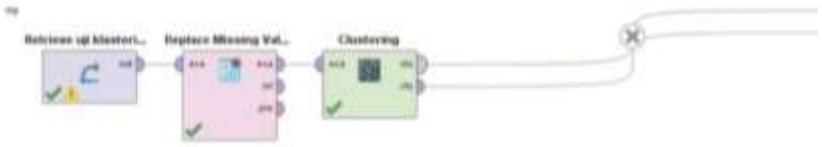
Tabel 1. Hasil Nilai Uji David Bouldin Index

Cluster	DBI
2	-1419.637
3	-1379.841
4	-1348.321
5	-1330.393
6	-1315.002
7	-1302.510
8	-1300.515
9	-1290.360
10	-1285.181
11	-1267.796
12	-1270.818
13	-1261.317
14	-1273.209
15	-1259.709

Dari hasil uji nilai David Bouldin Index yang telah dilakukan, maka hasil clustering optimal yang dapat digunakan dalam penelitian ini adalah K=15.

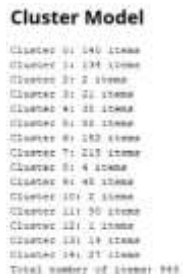
4. HASIL DAN PEMBAHASAN

Proses pengimplementasian K-Means dilakukan dengan menggunakan perangkat lunak Rapid Miner. Dari hasil penelitian dari 943 dataset hasil rating terhadap beberapa film yang telah di isi oleh berbagai macam jenis *age*, g*ender* dan *occopution*. Di dapati bahwa *cluster* di bagi menjadi 15 cluster sesuai degan uji Davies Bouldin index yang menyatakan bahwa 15 cluster merupakan angka terendah mendekati 0 yaitu -129,709 dengan hasil yang dapat dilihat pada Tabel 1. Berikut merupakan hasil konfigurasi data set kedalam algoritma K-Means.



Gambar 4. Konfigurasi Rapid Miner

Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2022
Institut Informatika dan Bisnis Darmajaya, 27 Agustus 2022
ISSN: 2598-0256, E-ISSN: 2598-0238



Gambar 5. Penempatan Data ke *Cluster*

Gambar 6. Rincian *Cluster*

5. KESIMPULAN

Dari permasalahan mengenai sistem rekomendasi yang mengalami kendala terkait adanya pengguna baru yang belum memberikan rating sehingga sulitnya untuk memberikan rekomendasi film-film terkait yang diduga disukai oleh pengguna tersebut, maka dapat di tarik kesimpulan sebagai berikut:

1. Mengelompokan data dengan menggunakan algoritma K-means dilakukan dengan cara menentukan jumlah cluster, hitung jarak terdekat dengan pusat cluster.
2. Klaster optimal yang dapat digunakan dalam algoritma K-means dalam permasalahan ini yaitu k=15
3. Ditemukan bahwa file clustering sebagai berikut
   Cluster 0; 140 items
   Cluster 1: 134 items
   Cluster 2: 2 items
   Cluster 3; 21 items
   Cluster 4: 33-items
   Cluster 5: 55 items
   Cluster 6: 152 items
   Cluster 7; 215 item3
   Cluster 8; 4 items
   Cluster 9: 48Items
   Cluster 10; 2 items
   Cluster 11: 90 1pens
   Cluster 12: 1 items
   Cluster 13; 14 items
   Cluster 14; 27 items

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih yang sebesar – besarnya kepada seluruh pihak yang telah membantu proses pelaksanaan penelitian ini.

DAFTAR PUSTAKA

Hu, Y. C. (2014). Recommendation using neighborhood methods with preference-relation-based similarity. *Information Sciences*, *284*, 18–30. https://doi.org/10.1016/j.ins.2014.06.043

Lien, D. T., & Phuong, N. D. (2014). Collaborative filtering with a graph-based similarity measure. *2014 International Conference on Computing, Management and Telecommunications, ComManTel 2014*, 251–256. https://doi.org/10.1109/ComManTel.2014.6825613

Huang, B. H., & Dai, B. R. (2015). A Weighted Distance Similarity Model to Improve the Accuracy of Collaborative Recommender System. *Proceedings - IEEE International Conference on Mobile Data Management*, *2*, 104–109. https://doi.org/10.1109/MDM.2015.43

Ruswanda, G. A. P., Baizal, Z. A., & Nasution, E. (2015). Penanganan Masalah Cold Start Dan Diversity Rekomendasi Menggunakan Item-Based Clustering Hybrid Method. *E- Proceeding of Engineering*, *2*(3), 8035–8041.

Al-Bashiri, H., Abdulgabber, M. A., Romli, A., & Hujainah, F. (2017). Collaborative filtering recommender system: Overview and challenges. *Advanced Science Letters*, *23*(9), 9045– 9049. https://doi.org/10.1166/asl.2017.10020

Tripathi, P., Vishwakarma, S. K., & Lala, A. (2016). Sentiment Analysis of English Tweets Using Rapid Miner. *Proceedings - 2015 International Conference on Computational Intelligence and Communication Networks, CICN 2015*, 668–672. https://doi.org/10.1109/CICN.2015.137

Laksana, E. A. (2014). Collaborative Filtering dan Aplikasinya. *Jurnal Ilmiah TeknologiInformasi Terapan*, *1*(2407–3911), 36–40.

Min, Z., & Shuzhen, Y. (2015). A collaborative filtering recommender algorithm based on the user interest model. *Proceedings - 17th IEEE International Conference on Computational Science and Engineering, CSE 2014, Jointly with 13th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2014, 13th International Symposium onPervasive Systems*, , 198–202. https://doi.org/10.1109/CSE.2014.67

Darwis, D., Siskawati, N., & Abidin, Z. (2021). Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional. *Jurnal Tekno Kompak*, *15*(1), 131. https://doi.org/10.33365/jtk.v15i1.744

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Computer Engineering, Science and System Journal*, *4*(1), 78. https://doi.org/10.24114/cess.v4i1.11458

Wulan Sari, R., Wanto, A., & Perdana Windarto, A. (2018). *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer) IMPLEMENTASI RAPIDMINER DENGAN METODE K-MEANS (STUDY KASUS: IMUNISASI CAMPAK PADA BALITA BERDASARKAN PROVINSI)*. *2*(1), 224–230. http://ejurnal.stmik-budidarma.ac.id/index.php/komik

E. Prasetyo, "Reduksi Dimensi Set Data dengan DRC pada Metode Klasifikasi SVM dengan Upaya Penambahan Komponen Ketiga", Prosiding SNATIF, 2014, 293-300.