Gunung Djati Conference Series, Volume 3 (2021)
Mini Seminar Kelas *Data Mining* 2020
ISSN: 2774-6585
Website: https://conferences.uinsgd.ac.id/gdcs

Analisis Pendapatan dan Pengeluaran Film menggunakan Algoritma *Bisecting K-Means*

(*Analysis of Film Budget and Profit using the Bisecting K-Means Algorithm*)

Ahmad Fauzi[1], Deden Muhamad Furqon[2], Riki Ahmad Maulana[3], Nurul Dwi Cahya[4], Muhammad Nur Sidiq[5]

[1]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1177050009@student.uinsgd.ac.id
[2]Teknik Informatika, UIN Sunan Gunung Djati Bandung, furqoncreative24@gmail.com
[3]Teknik Informatika, UIN Sunan Gunung Djati Bandung, ahmadriki9512@gmail.com
[4]Teknik Informatika, UIN Sunan Gunung Djati Bandung, nuruldwicahya925@gmail.com
[5]Teknik Informatika, UIN Sunan Gunung Djati Bandung, 1167050104@student.uinsgd.ac.id

Abstrak

Seiring perkembangan film yang semakin kompetitif pada beberapa elati terakhir ini, menjadikan ekosistem perfilman perlu mendapat perhatian lebih bagi para stakeholder yang berkecimpung di dalamnya untuk terus melakukan berbagai tindakan inovasi dan menciptakan strategi pemasaran ekonomi kreatif yang lebih efektif. Dengan memanfaatkan dataset yang ada, para produsen konten perfilman dapat membangun suatu elati rekomendasi yang dapat menunjang proses assessment business model dan analisa konsep produk kreatif perfilman yang akan diluncurkan di pangsa pasar perfilman yang ada sehingga kelak dapat dihasilkan perencanaan dan perancangan konsep produksi film yang lebih *profitable* dan *sustainable* dari segi pendanaan (*budget*) dan proyeksi pendapatan (*profit gross*). Penelitian ini merupakan suatu bentuk elaborasi terhadap perancangan elati rekomendasi dengan menggunakan Algoritma Bisecting K-Means untuk dapat menghasilkan suatu hasil analisa berupa klasifikasi (*classification*) terhadap berbagai produk film yang terdapat pada dataset yang sudah dihimpun sebanyak 5048 baris data dengan mengambil porsi data sebanyak 1000 baris sebagai alokasi data khusus untuk melakukan training pada elati. Data yang telah dialokasikan selanjutnya akan dilakukan proses klasterisasi dengan membagi ke dalam 4 (empat) buah cluster berbeda yang masing-masing merupakan hasil regresi berdasarkan parameter yang telah ditetapkan sebelumnya dan membentuk sebuah centroids yang merupakan nilai rata-rata dari seluruh node cluster yang dibangun.

Kata kunci: classification, bisecting k-means, budget, film, profit

Abstract

As the development of the film industry has become increasingly competitive in the last few decades, the film ecosystem needs to get more attention for stakeholders involved in it to continue to carry out various innovative actions and create more effective creative economy marketing strategies. By utilizing existing datasets, film content producers can build a recommendation system that can support the process of assessing business models and analyzing the concept of creative film products that will be launched in the existing film market so that later planning and design of more profitable film production concepts can be produced. And sustainability in terms of funding (budget) and projected revenue (gross profit). This research is a form of elaboration on the design of a

recommendation system using the Bisecting K-Means Algorithm to be able to produce an analysis result in the form of classification of various film products contained in a dataset that has been collected as many as 5048 rows of data by taking 1000 lines of data. As a special data allocation for conducting training on the system. The data that has been allocated will then be carried out by the clustering process by dividing into 4 (four) different clusters, each of which is the result of regression based on predetermined parameters and forming a centroids which is the average value of all cluster nodes built.

Keywords: *classification, bisecting k-means, budget, film, profit*

1 Pendahuluan

Saat ini, media komunikasi massa terus menunjukkan perkembangan yang sangat pesat, tak terkecuali media komunikasi berbasis Film. Telah diketahui oleh banyak kalangan masyarakat bahwa film merupakan komoditas utama dalam memenuhi kebutuhan manusia dalam melakukan rekreasi. Hal tersebut menjadi elati utama yang mendasari terbentuknya elative film yang elati dan menjadi pemicu utama bagi para stakeholder terkait yang berkepentingan untuk dapat terus melakukan inovasi terhadap keberadaan konten hiburan berbasis film. Perkembangan dan inovasi perfilman tentunya memberikan dampak signifikan terhadap peningkatan kebutuhan pangsa pasar akan jumlah film dengan genre yang makin bervariasi dan membangkitkan gaya hidup baru di tengah masyarakat dimana aktivitas menonton film merupakan suatu kebutuhan tersendiri dalam menjalani kehidupan sehari-hari.

Inovasi yang tercipta di dalam elative perfilman menuntut para stakeholder yang turut andil di dalamnya untuk dapat menciptakan layanan perfilman yang menarik dan intuitif bagi para penikmat film itu sendiri sehingga secara simultan dapat meningkatkan daya pikat (engagement) dan memperbesar tingkat retensi pengguna terhadap layanan yang dihadirkan. Untuk dapat mengakomodasi potensi tersebut, maka perlu hadir suatu elati rekomendasi berbasis data yang dapat memberikan panduan bagi pengelola layanan dalam mengembangkan business model nya agar dapat terus catch-up dengan perubahan demand di tengah masyarakat serta menjadi tools utama bagi pengguna layanan untuk dapat terus menikmati konten film sesuai dengan preferensi yang diinginkan. Untuk itu, maka dengan penelitian ini kami mencoba untuk melakukan elaborasi lebih lanjut terkait dengan penggunaan elati rekomendasi yang mengimplementasikan metode clustering dengan memanfaatkan efektivitas kinerja dari Algoritma Bisecting K-Means yang merupakan algoritma yang memiliki performa lebih baik dibandingkan algoritma K-Means karena memproduksi cluster yang seragam dan tidak memproduksi cluster kosong, sehingga dapat memberikan tingkat keakurasian yang lebih presisi dan diketahui memiliki efisiensi yang tinggi ketika jumlah cluster meningkat [1].

Sebagaimana penelitian yang telah dilakukan sebelumnya oleh Arwin Halim, dkk. (2017) dimana telah diketahui penggunaan Algoritma Bisecting K-Means yang ditunjang dengan metode Collaborative Learning dapat menghasilkan performa elati rekomendasi yang handal dan memiliki tingkat kesalahan (error rate) yang elative kecil. Secara teknis, algoritma ini bekerja dengan mengelompokan data ke dalam setiap klaster (clustering) yang sesuai dengan titik pusat (centroid) dari masing masing klister [2], [3].

2 Metodologi

2.1 Data Mining

Data mining merupakan proses pengekstrakan informasi dari jumlah kumpulan data yang besar dengan menggunakan algoritma dan tehnik gambar dari statistik, mesin pembelajaran dan sistem

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

manajemen database [4]. Data mining yang disebut juga dengan Knowledge Discovery in Database (KDD) adalah sebuah proses secara otomatis atas pencarian data di dalam sebuah memori yang amat besar dari data untuk mengetahui pola dengan menggunakan alat seperti klasifikasi, hubungan (association) atau pengelompokan (clustering).

Secara sederhana, data mining dapat diartikan sebagai proses mengekstrak atau "menggali" pengetahuan yang ada pada sekumpulan data. Banyak orang yang setuju bahwa data mining adalah sinonim dari Knowledge Discovery in Database atau yang biasa disebut KDD. Dari sudut pandang yang lain, data mining dianggap sebagai satu langkah yang penting didalam proses KDD.

Menurut Han, J. And Kamber, M, 2001, proses KDD ini terdiri dari langkah-langkah sebagai berikut [5]:

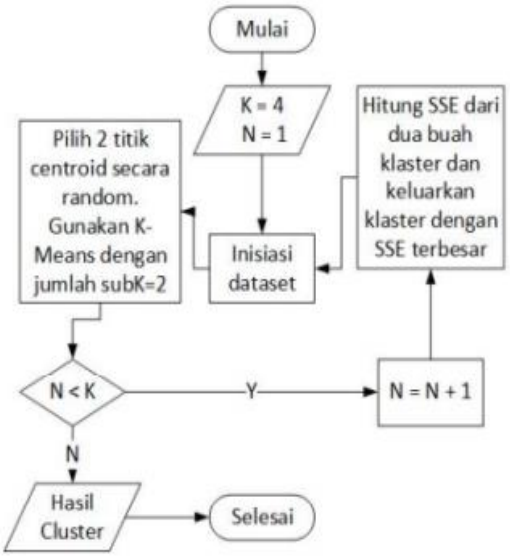
1. Data Cleaning, proses menghapus data yang tidak konsisten dan kotor
2. Data Integration, penggabungan beberapa sumber data
3. Data Selection, pengambilan data yang akan dipakai dari sumber data
4. Data Transformation, proses dimana data ditransformasikan menjadi bentuk yang sesuai untuk diproses dalam data mining
5. Data Mining, suatu proses yang penting dengan melibatkan metode untuk menghasilkan suatu pola data
6. Pattern Evaluation, proses untuk menguji kebenaran dari pola data yang mewakili knowledge yang ada didalam data itu sendiri
7. Knowledge Presentation, proses visualisasi dan teknik menyajikan knowledge digunakan untuk menampilkan knowledge hasil mining kepada user

2.2 Clustering

Proses pengelompokan sekumpulan obyek kedalam kelas-kelas obyek yang sama disebut clustering. Pengklasteran merupakan satu dari sekian banyak fungsi proses data mining untuk menemukan kelompok atau identifikasi kelompok obyek yang hampir sama. Analisis kluster (Clustering) merupakan usaha untuk mengidentifikasi kelompok obyek yang mirip-mirip dan membantu menemukan pola penyebaran dan pola hubungan dalam sekumpulan data yang besar. Hal penting dalam proses pengklasteran adalah menyatakan sekumpulan pola ke kelompok yang sesuai yang berguna untuk menemukan kesamaan dan perbedaan sehingga dapat menghasilkan kesimpulan yang berharga.

2.3 Bisecting K-means

Bisecting K-Means adalah variasi dari algoritma K-Means [6]. Kunci dari algoritma ini adalah satu cluster dibagi menjadi dua sub-cluster di setiap langkah. Algoritma ini mirip dengan algoritma pengelompokan hierarki. Sebenarnya, algoritma pengelompokan hierarki memiliki keuntungan karena tidak memerlukan jumlah kluster secara apriori, karena kluster dibagi dua di setiap langkah. Namun, dalam algoritma ini, masalahnya adalah dalam menentukan aturan berhenti, yaitu memutuskan apakah dan cluster mana yang harus tetap dibagi dua. Untuk tujuan ini, dua pendekatan utama digunakan: yang pertama menerapkan strategi sederhana membagi dua cluster terbesar dan yang kedua adalah membagi cluster dengan varians terbesar sehubungan dengan sentroid cluster [7]. Gambar 1 menunukkan diagram alir algoritma Bisecting K-Means.



Gambar 1 Diagram Alir Algoritma Bisecting K-Means

Untuk setiap cluster, kumpulkan semua synsets beserta frekuensi termnya, untuk semua dokumen yang ada di cluster tertentu itu. Atur semua synsets sesuai dengan urutan frekuensi istilah untuk synsets yang dikumpulkan. Sekarang beri label cluster dengan synset paling atas. Ini memberikan gambaran tentang jenis dokumen yang terdapat dalam cluster tertentu. Ulangi prosedur yang sama untuk semua cluster.

Untuk pengelompokan, dua ukuran kebaikan cluster atau kualitas cluster digunakan. Jenis ukuran pertama menentukan untuk membandingkan kumpulan cluster yang berbeda tanpa mengacu pada pengetahuan eksternal dan dinamai sebagai ukuran kualitas internal. Jenis ukuran kedua menentukan untuk memungkinkan kita menghitung seberapa baik pengelompokan bekerja dengan membandingkan kelompok yang dihasilkan oleh teknik pengelompokan dengan kelas yang ditentukan. Jenis ukuran ini disebut sebagai ukuran kualitas eksternal. Salah satu ukuran kualitas internal untuk memeriksa kualitas cluster adalah RMSE (Root Mean Square Error), dapat dihitung seperti yang ditunjukkan pada Persamaan [8].

3 Hasil dan Pembahasan

Algoritma bisecting k-means digunakan dalam penelitian ini karena mampu menginisialisasi *centroid* secara acak dan melakukan proses *bisecting* (pembagian menjadi dna) pada klaster dengan mat *sum square error* (SSE) maksimum atau klaster terbesar.. Tahapan dari algoritme bisecting k-means secara berurutan adalah menentukan klaster yang akan dipisah *[split j,* menemukan 2 sub—klaster menggunakan k-means tipe dasar (tahap *bisecting),* membagi dna ITER waktu dan ambit hasil *split clustering* yang memiliki SSE tertinggi, dan men gulangi langkah- langkah sebeluinnya hingga jumlah Master tercapai.

Calculate Centroids

Menghtitung nilai centroids ini digunakan untuk menentukan titik berkumpulnya clustering dimana hasil dari perhitungan akan dikumpulkan pada sebuah diagram dan mencari kecocokan Antara nilai yang dibandingkan dengan nilai centroids [9]. Semakin kecil nilai kecocokan, berarti semakin identic nilai tersebut dengan centroids yang diinginkan, setelah bisecting dilakukan. Maka system akan

Gunung Djati Converence Series
ISSN 2774-6585
Volume 3 (2021)

menghitung nilai centroids yang dibagi kedalam n clusters, pada penilitian ini kami menggunakan sebanyak 4 buah clusters centroids pada nilai tengah.

Hitung hasil SSE, Inter Class, dan Intra Class

Pada tahap ini selanjutnya saatnya untuk menghitung nilai SSE / Sum squared errors untuk mengetahui jarak tiap clusters dengan pusat cluster. Jika semua clusters identic, maka SSE akan bernilai 0, rumus untuk menghitung SSE adalah :

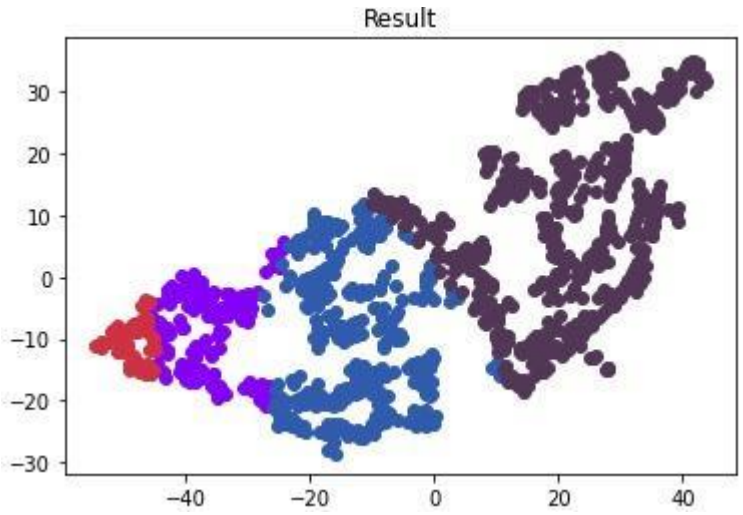
$$\sqrt{((x1-y1)^2 + (x2-y2)^2} \qquad (1)$$

SSE digunakan agar setiap clusters mudah untuk di observasi, setelah nilai SSE didapatkan maka mengobservasi data. Setelah nilai SSE didapatkan saatnya untuk menghitung nilai intra dimana intra digunakan untuk mengetahui nilai data dari inti cluster, dan inter untuk mengetahui data dengan cluster terdekat.

Hasil Visualisasi

Pada tahap ini setiap kemudian seluruh nilai pada setiap kalkulasi tersebut di implementasikan kedalam sebuah graphic agar memudahkan kita untuk memvisualisasikan tiap-tiap cluster nya berdasarkan data yang telah di hitung. Pada visualisasi ini menggunakan library matploitlib pada python dan kemudian ditampilkan dalam sebuah graphic secara otomatis dengan cara menginputkan parameter nilai yang telah kita kalkulasi, yaitu budget, dan keuntungan tiap'' film.

Pada hasil dan pembahasan ini kami telah berhasil membagi data film yang di input menjadi 4 buah cluster dengan menunjukan nilai SSE, Inter Class, dan juga Intra Class. Data yang di oleh berjumlah 1000 buah dan kami berhasil menempatkan nya menjadi 4 cluster. Titik centroids yang digunakan adalah nilai tengah dari data, karena data berjumlah 1000 / 2, maka titik yang digunakan sebagai pusat clustering adalah data ke 500 dan seterusnya. Setelah semua perhitungan dilakukan, maka didapatkan lah visualisasi pada Gambar 2.



Gambar 2 Visualisasi hasil klasterisasi

Intra =[80361275.5854854, 63479003.68008132, 41134549.78689848, 140073280.4665961] 81262027.37976533 1346506308326928.5 36694772.220671006

```
Inter =[132321800.12245046, 202897224.0564931, 199326231.69519284,
95714926.50755188, 302009986.62452275, 377004495.38328105]
218212444.06491533 9201838867845366.0 95926215.74859172
SSE   =[5.427654240428392e+17, 8.65688230138127e+17,
5.803761493607855e+17, 7.150141192380417e+17] 6.759609806949484e+17
1.6099773276096438e+34 1.2688488198401115e+17
```

Setiap cluster memiliki warna yang berbeda dan juga dideskripsikan berdasarkan nilai intra, inter,dan juga SSE yang berada dibawahnya. Hal ini bisa membantu setiap perusahaan untuk melakukan klasifikasi pada film mereka kedepannya dan berfungsi sebagai analisis penjualan/perilaku pasar.

4 Simpulan

Pada zaman sekarang analisis tidak harus digunakan secara manual dan membuang banyak waktu, di era 4.0 teknologi ini kita mampu melakukan analisis yang cepat untuk menentukan mana dari produk kita yang mengalami kerugian, produk unggulan, atau produk yang memiliki potensi di masa depan dengan cara membaca data. Hal ini dilakukan untuk menopang sebuah bisnis dalam rangka menganalisis potensi dan keinginan customer akan sebuah produk. Dari data clustering tersebut, selanjutnya kita bisa menentukan langkah yang tepat bagi perusahaan untuk kedepannya. Hasil yang ditunjukan pada grafik menunjukan bahwa data budget dan gross yang berada pada dataset sudah dibagi kedalam 4 cluster yang berbeda, dimana X menunjukan budget rata-rata dan Y menunjukan gross/penghasilan rata rata (kotor) pada film tersebut. Dibawah grafik terdapat nilai SSE, Inter class, dan juga intra class. Dimana nilai inter adalah nilai jarak rata-rata node pada titik sebuah cluster pusat. Intra adalah nilai dari rata-rata node ke cluster terdekat. Sedangkan SSE menunjukan nilai rata-rata kemiripan sebuah cluster terhadap centroids

Referensi

[1] A. Halim, H. Gohzali, D. M. Panjaitan, and I. Maulana, “Sistem Rekomendasi Film menggunakan Bisecting K-Means dan Collaborative Filtering,” *Citisee*, vol. 1, no. 3, pp. 37–41, 2017.

[2] N. Puspitasari, J. A. Widians, and N. B. Setiawan, “Customer segmentation using bisecting k-means algorithm based on recency, frequency, and monetary (RFM) model,” *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 78–83, 2020, doi: 10.14710/jtsiskom.8.2.2020.78-83.

[3] N. Puspitasari, J. A. Widians, and N. B. Setiawan, “Segmentasi pelanggan menggunakan algoritme bisecting k-means berdasarkan model recency, frequency, dan monetary (RFM),” *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 78–83, 2020.

[4] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction do Data Mining*. 2005.

[5] H. Jiawei, M. Kamber, J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2006.

[6] R. Patil and A. Khan, “Bisecting K-Means for Clustering Web Log data,” *Int. J. Comput. Appl.*, vol. 116, no. 19, pp. 36–41, 2015, doi: 10.5120/20448-2799.

[7] B. S. V. Krishna, P. Satheesh, and R. Suneel Kumar, “Comparative study of k-means and bisecting k-means techniques in wordnet based document clustering,” *Int. J. Eng. Adv. Technol.*, vol. 1, no. 6, pp. 1–4, 2012.

[8] F. Ridho and A. A. Kusuma, “Deteksi Intrusi Jaringan dengan K-Means Clustering pada Akses Log dengan Teknik Pengolahan Big Data,” *J. Apl. Stat. Komputasi Stat.*, vol. 10, no. 1, p. 53, 2019, doi: 10.34123/jurnalasks.v10i1.202.

[9] D. Herawatie, E. Wuryanto, and P. Purbandini, “Perbandingan Algoritma Pengelompokan Non-Hierarki untuk Dataset Dokumen,” in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2014, vol. 1, no. 1.