 ISSN (Online) 2828-786X

PENGGUNAAN ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN REKOMENDASI FILM INDONESIA

USING K-MEANS CLUSTERING ALGORITHM TO DETERMINE INDONESIAN FILM RECOMMENDATIONS

Elsa Vania[1)], Salma Nuraini[2)], Dhian Satria Yudha Kartika[3)]
E-mail : [1)]19082010027@student.upnjatim.ac.id , [2)]19082010075@student.upnjatim.ac.id ,
[3)]dhian.satria@upnjatim.ac.id

[1,2,3] Sistem Informasi, Fakultas Ilmu Komputer, UPN "Veteran" Jawa Timur

Abstrak

Seiring berkembangnya industri film, semakin banyak pula film yang diproduksi. Banyaknya film ini membuat penonton bimbang untuk memilih film mana yang akan ditonton. Penggunaan algoritma *k-means clustering* dapat membantu dalam mengelompokkan film berdasarkan karakteristiknya, sehingga penonton dapat memilah film dengan mudah. Tahapan klasterisasi dilakukan dengan metode CRISP-DM. Sedangkan algoritma yang diterapkan adalah K-*Means*. Dataset yang digunakan diambil dari kaggle yang berisi data film indonesia hasil scraping dari website IMDB dengan data film sebanyak 1272. Hasil dari tahapan klasterisasi ditemukan bahwa ada dua kelompok film yaitu film yang direkomendasikan dan film yang kurang direkomendasikan. Dari hasil klaster tersebut dapat menghasilkan rekomendasi film Indonesia yang mungkin bisa menjadi referensi untuk ditonton.
Kata kunci: *K-Means*, *Clustering*, CRISP-DM, Film Indonesia

Abstract

As the film industry develops, more and more films are produced. The number of these films makes the audience hesitate to choose which film to watch. The use of the k-means clustering algorithm can help in grouping films based on their characteristics, so that viewers can sort films easily. The clustering stage was carried out using the CRISP-DM method. While what is applied is K-Means. The dataset used is taken from Kaggle which contains Indonesian film data scraped from the IMDB website with 1272 film data. The results of the clustering stage found that there are two groups of films, namely recommended films and films that are less certain. From the results of the cluster, it can produce recommendations for Indonesian films that might be a reference to watch.
***Keywords*:** *K-Means, Clustering, CRISP-DM, Indonesian Movie*

1. PENDAHULUAN

Film merupakan media audio visual yang berisikan serangkaian gambar bergerak yang ditampilkan pada layar [1]. Perkembangan film memiliki perjalanan cukup panjang dari tahun ke tahun hingga sekarang menjadi film yang penuh dengan efek dan sangat mudah dicari sebagai media hiburan. Di Indonesia sendiri, film pertama diciptakan pada tahun 1900 dan mulai berkembang sekitar tahun 1980-an hingga sekarang [2].

Sampai sekarang terdapat banyak judul film dari berbagai negara, salah satunya Indonesia. Banyaknya film ini membuat penonton bingung untuk memilih film yang akan ditonton. Ada banyak faktor yang mempengaruhi penonton untuk menilai sebuah film, mulai dari genre, aktor yang memerankan, sutradara, durasi, penilaian dari penonton lain, dll. Dengan melakukan pengelompokan film dapat mempermudah penonton untuk memilah film mana yang akan ditonton.

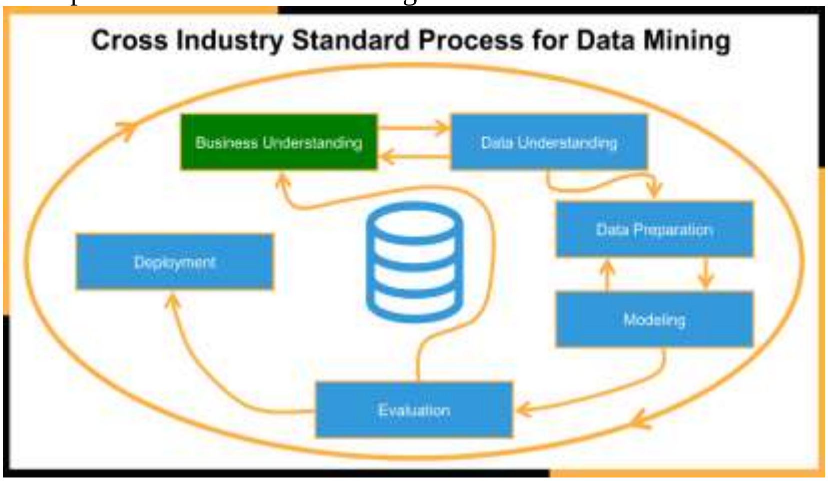
Penelitian ini bertujuan untuk mengelompokkan film produksi Indonesia untuk menentukan rekomendasi film Indonesia. Dalam penelitian ini *k-means clustering* digunakan

 ISSN (Online) 2828-786X

untuk mengelompokkan film menjadi dua kelompok, yaitu direkomendasikan dan kurang direkomendasikan. *Clustering* merupakan salah satu metode *data mining* yang mengelompokkan suatu data yang memiliki kesamaan karakteristik ke dalam kelompok-kelompok atau klaster [3]. Terdapat 2 jenis *clustering* yang dipakai untuk mengelompokkan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering*. *K-Means Clustering* merupakan metode *non-hierarchical clustering* yang mengelompokan data dalam bentuk klaster atau kelompok [4]. Pengelompokan ini dilakukan berdasarkan rating atau penilaian dari penonton dan jumlah penonton yang memberikan rating.

2. METODOLOGI

Metodologi penelitian ini mengacu pada metode CRISP-DM. Metode ini memiliki enam tahapan yang berurutan dari pemahaman bisnis hingga *deployment*. Setiap tahapan pada metode ini bersifat adaptif dimana setiap tahapan selanjutnya bergantung atas tahapan sebelumnya [5]. Adapun untuk tahapan CRISP-DM adalah sebagai berikut :



Gambar 1. Tahapan CRISP-DM [6]

2.1 Business Understanding

Pada tahapan ini, peneliti melakukan pemahaman dari tujuan di lingkup bisnis di suatu proyek. Adapun tujuan dari penelitian ini guna mengelompokkan (*clustering*) dari data film indonesia guna mendapatkan rekomendasi tontonan film Indonesia yang mengacu pada penilaian penonton. Tujuan serta batasan yang diterjemahkan menjadi formula suatu permasalahan dalam kasus *data mining* [5].

2.2 Data Understanding

Dataset yang digunakan pada penelitian ini merupakan kumpulan film Indonesia yang bersumber dari situs kaggle. Atribut data yang akan digunakan dalam *clustering* adalah dengan data *users rating* dan *votes*. Total data yang digunakan sebanyak 1272 judul film.

2.3 Data Preparation

Untuk tahapan ini, data yang sudah didapat, dilakukan pengecekan apakah atribut yang digunakan bernilai *null* (kosong) atau tidak. Untuk data yang digunakan tidak terdapat data yang kosong sehingga tidak ada penghapusan data yang bernilai *null* atau kosong.

2.4 Modeling

Mengaplikasikan model yang tepat guna memperoleh hasil yang baik. Untuk melakukan pengklasteran ini, digunakan algoritma *K-Means*. Adapun langkah yang dilakukan untuk menerapkan algoritma ini adalah [7] :

1. Menentukan jumlah K *cluster*.
2. Menentukan pusat *cluster* secara acak menurut K-*Cluster* yang dibentuk di awal.
3. Menghitung jarak cluster dengan centroid dengan rumus *Euclidean Distance*.

$$d(x_i, \mu_i) = \sqrt{(x_i - \mu_i)^2} \quad (1)$$

Untuk:
xi = bobot kata ke i pada *cluster* yang ingin dicari jaraknya
μj = μi bobot kata ke i pada pusat *cluster*.
4. Mengelompokan objek (data) menurut jarak minimum dengan centroid.
5. Memperbarui nilai centroid dari rata-rata *cluster* dengan rumus:

$$C_k = \frac{1}{n_k}\sum \quad d_i \quad \text{(2)}$$

Untuk :
nk = jumlah data dalam *cluster*
di = jumlah dari nilai jarak yang masuk dalam masing-masing *cluster*
6. Mengulangi langkah 2 hingga langkah 5 sampai tidak ditemukan perubahan cluster.

2.5 Evaluasi

Tahap ini melakukan evaluasi atau menilai atas penetapan model yang ditentukan. Pada penelitian ini tidak melakukan tahap ini dikarenakan hanya melakukan *clustering* dengan tujuan mengelompokkan data.

2.6 Deployment

Melakukan implementasi model yang telah ditentukan. Hal ini dilakukan dengan menentukan tiap instance film sesuai dengan klasternya. Setiap film dalam suatu klaster memiliki atribut *users rating* dan votes yang memiliki kemiripan.

3. HASIL DAN PEMBAHASAN

Dataset yang digunakan pada penelitian ini adalah data film Indonesia yang didapatkan dari website IMDB sebanyak 1272 judul film yang terdiri dari *data movie_id, title, year, description, genre, rating, users_rating, votes, languages, directors, actors, runtime*. Kemudian data film akan diolah dengan algoritma *k-means clustering* berdasarkan users_rating dan votes.

3.1 Data film Indonesia

	movie_id	title	year	description	genre	rating	users_rating	votes	languages	directors	actors	runtime
0	100001	#FriendButMarried 2	2020	Ayudia (Mawar De Jongh) is not satisfied enoug...	Biography	13+	6.5	120	Indonesian	Rako Prijanto	['Adipati Dolken', 'Mawar Eva de Jongh', 'Vonn...	100 min
1	100002	4 Mantan	2020	Sara, Airin, Rachel, and Amara were accidental...	Thriller	17+	6.4	8	Indonesian	Hanny Saputra	['Ranty Maria', 'Jeff Smith', 'Melanie Berentz...	80 min
2	100003	Aku Tahu Kapan Kamu Mati	2020	After apparent death, Siena is able to see sig...	Horror	13+	5.4	17	Indonesian	Hadrah Daeng Ratu	['Natasha Wilona', 'Ria Ricis', 'Al Ghazali', ...	92 min
3	100004	Anak Garuda	2020	Good Morning Indonesia, a school for poor orph...	Adventure	13+	9.1	27	Indonesian	Faozan Rizal	['Tissa Biani Azzahra', 'Violla Georgie', 'Aji...	129 min
4	100005	Dignitate	2020	Alfi (Al Ghazali) meets Alana (Caitlin Haldem...	Drama	17+	7.6	33	Indonesian	Fajar Nugros	['Al Ghazali', 'Caitlin Halderman', 'Giorgino ...	109 min

Gambar 2. Data Film Indonesia

Penjelasan maksud setiap kolom:

movie_id	: ID dari setiap film
title	: Judul film
year	: Tahun perilisan film
description	: Sinopsis film
genre	: Genre film
rating	: Batasan umur film
users_rating	: Penilaian penonton terhadap film
votes	: Jumlah penonton yang memberikan penilaian terhadap film
languages	: Bahasa pada film
directors	: Sutradara film
actors	: Aktor yang memerankan
runtime	: Durasi waktu film

 ISSN (Online) 2828-786X

Sebelum data diolah, dilakukan eksplorasi data terlebih dahulu untuk melihat tipe data tiap atribut, deskriptif statistik data, dan mengecek adanya data yang bernilai null.

movie_id	int64
title	object
year	int64
description	object
genre	object
rating	object
users_rating	float64
votes	int64
languages	object
directors	object
actors	object
runtime	object
dtype: object	

Gambar 3. Tipe Data tiap Kolom

	movie_id	year	users_rating	votes
count	1272.00000	1272.000000	1272.000000	1272.000000
mean	100636.50000	2007.023585	6.144418	86.988208
std	367.33908	12.968560	1.389315	151.539191
min	100001.00000	1926.000000	1.200000	5.000000
25%	100318.75000	2006.000000	5.300000	12.000000
50%	100636.50000	2011.000000	6.400000	27.000000
75%	100954.25000	2016.000000	7.100000	76.000000
max	101272.00000	2020.000000	9.400000	991.000000

Gambar 4. Deskriptif Statistik Data

Data Null?	
movie_id	0
title	0
year	0
description	432
genre	36
rating	896
users_rating	0
votes	0
languages	0
directors	7
actors	0

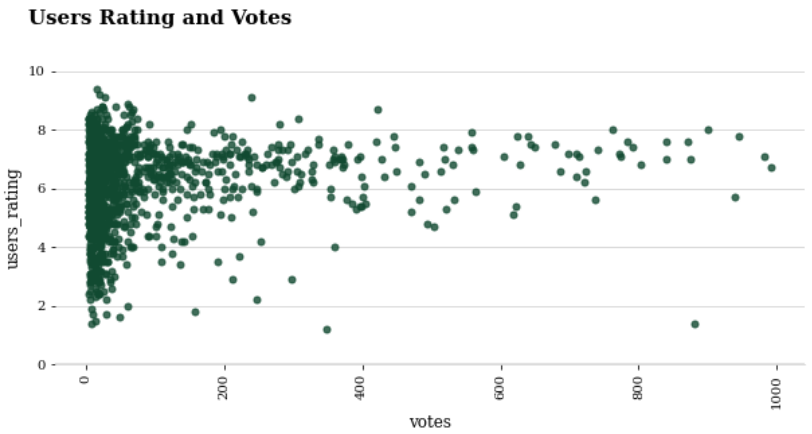
Gambar 5. Jumlah Data Null tiap Kolom

Data yang akan digunakan adalah data users_rating dan votes. Dapat dilihat pada gambar diatas data users_rating dan votes memiliki tipe data float dan integer, sehingga tidak diperlukan pengubahan tipe data. Kedua data ini juga tidak memiliki nilai *null*, sehingga tidak diperlukan pengisian *missing value*.

3.2 Visualisasi data users_rating dan votes

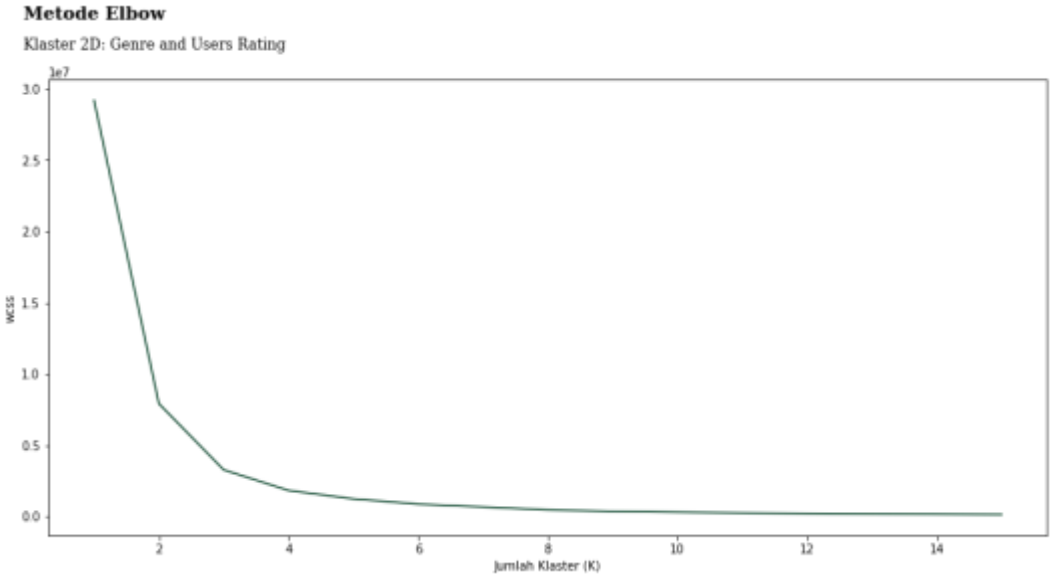
Visualisasi dilakukan untuk mengecek adanya klaster diantara atribut. Visualisasi ditampilkan dalam bentuk grafik 2D.

 ISSN (Online) 2828-786X



Gambar 6. Visualisasi users_rating dan votes

3.3 Jumlah klaster optimal



Gambar 7. Grafik Metode *Elbow*

Dalam menentukan jumlah klaster optimal digunakan metode *elbow*. Pada metode *elbow* jika nilai klaster pertama dan kedua mengalami penurunan grafik yang besar, maka jumlah klaster tersebut yang paling optimal [8]. Pada gambar grafik di atas dapat dilihat penurunan drastis terdapat pada angka 2. Sehingga dapat disimpulkan jumlah klaster optimal adalah 2.

3.4 Validasi klaster

Validasi klaster dilakukan untuk mengetahui seberapa baik klaster yang terbentuk dari dataset untuk dilakukan observasi. Dalam melakukan validasi klaster, metode yang digunakan adalah metode *silhouette coefficient*. *Silhouette coefficient* merupakan gabungan metode *cohessian* dan *separation* [9]. Pengukuran *cohessian* dilakukan dengan perhitungan seluruh objek dalam sebuah cluster, sedangkan pengukuran *separation* dilakukan dengan perhitungan jarak rata-rata setiap objek dalam sebuah klaster dengan klaster terdekatnya [10]. *Silhouette coefficient* bernilai antara -1 hingga 1. Pengelompokan data pada satu klaster dikatakan baik jika nilainya mendekati angka 1.

```
Jumlah Klaster = 2 Nilai Rata-Rata Silhouette = 0.8329264022333984
Jumlah Klaster = 3 Nilai Rata-Rata Silhouette = 0.7951753590867102
Jumlah Klaster = 4 Nilai Rata-Rata Silhouette = 0.7496990471629705
Jumlah Klaster = 5 Nilai Rata-Rata Silhouette = 0.7077705635161482
Jumlah Klaster = 6 Nilai Rata-Rata Silhouette = 0.6968612585434968
```

Gambar 8. Perhitungan Silhouette Coefficient

 ISSN (Online) 2828-786X

Dari gambar di atas, dapat dilihat hasil nilai rata-rata *silhouette* terbesar berada pada klaster dengan nilai K = 2. Hasil perhitungan ini sesuai dengan jumlah K yang ditentukan pada metode *elbow*.

3.5 Hasil Klaster

Tahap selanjutnya adalah menghitung nilai centroid dan mengelompokkan film berdasarkan klasternya. Setelah perhitungan *k-means clustering* didapatkan hasil sebagai berikut.

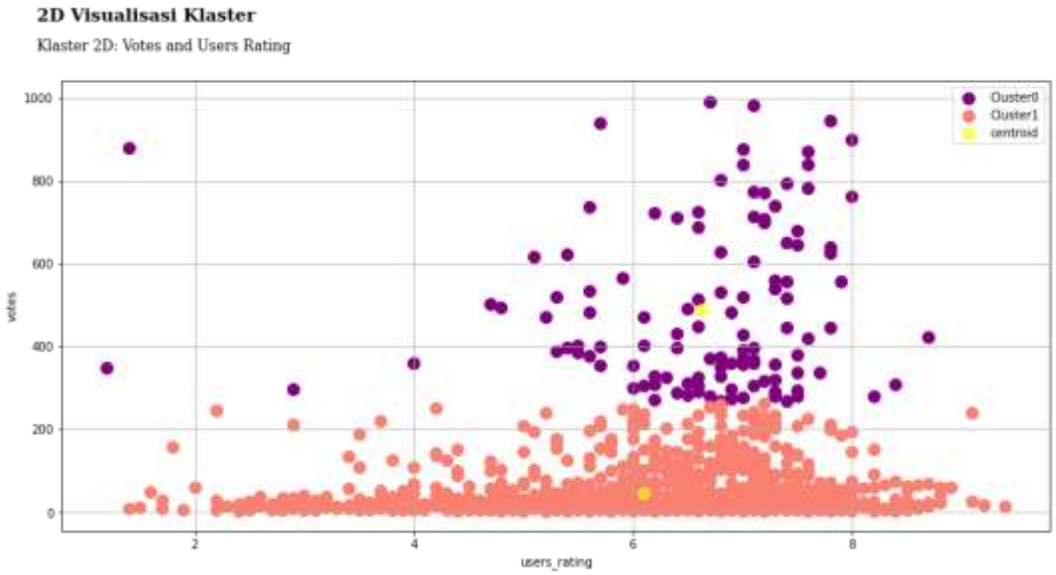
	title	users_rating	votes	Cluster
0	#FriendButMarried 2	6.5	120	1
1	4 Mantan	6.4	8	1
2	Aku Tahu Kapan Kamu Mati	5.4	17	1
3	Anak Garuda	9.1	27	1
4	Dignitate	7.6	33	1
...	...	...	...	...
1267	The Tiger from Tjampa	6.4	30	1
1268	Enam Djam di Djogja	6.3	9	1
1269	Darah dan Doa	6.6	27	1
1270	Resia Boroboedoer	7.0	8	1
1271	Loetoeng Kasaroeng	7.2	11	1

Gambar 9. Hasil Clustering

	Cluster	users_rating	votes
0	0	6.6	489.5
1	1	6.1	45.4

Gambar 10. Centroid tiap Klaster

Dapat dilihat pada gambar diatas terdapat 2 klaster, yaitu klaster 0 dan 1. Klaster 0 memiliki rata-rata users_rating dan votes yang lebih tinggi dari klaster 1. Sehingga dapat disimpulkan klaster 0 merupakan kelompok film yang direkomendasikan dan klaster 1 merupakan kelompok film yang kurang direkomendasikan.



Gambar 11. Visualisasi Klaster

Gambar di atas menunjukkan visualisasi hasil clustering berdasarkan users_rating dan votes. Klaster 0 memiliki users_rating dan votes yang tinggi. Sehingga film dalam klaster 0 merupakan kelompok film yang direkomendasikan. Sedangkan klaster 1 memiliki jumlah

users_rating dan votes yang lebih sedikit. Sehingga film dalam klaster 1 dapat dikelompokkan sebagai film yang kurang direkomendasikan. Berikut adalah beberapa judul film sesuai klasternya.

Klaster 0 (direkomendasikan)
1. A Man Called Ahok
2. Yowis Ben
3. 27 Steps of May
4. Dua Garis Biru
5. The Raid 2
6. Keluarga Cemara
7. Imperfect
8. Laskar Pelangi
9. What's Up With Love?
10. Petualang Sherina
11. dst

Klaster 1 (kurang direkomendasikan)
1. Skandal Cinta Babi Ngepet
2. Legend of the East
3. Genderuwo
4. Roy Kiyoshi: The Untold Story
5. Triangle the Dark Side
6. Malam Jumat Kliwon
7. Dendam dari Kuburan
8. Kuntilanak Kesurupan
9. Pengantin Pantai Biru
10. The Empire's Throne
11. Dst

4. KESIMPULAN DAN SARAN

Penelitian ini menghasilkan klaster film Indonesia yang terbagi dalam dua klaster yaitu film yang direkomendasikan untuk ditonton dan film yang kurang direkomendasikan untuk ditonton berdasarkan jumlah penonton yang memberi rating dan penilaian (rating) dari penonton itu sendiri. Untuk film yang direkomendasikan untuk ditonton berada di Cluster 0 dengan jumlah 119 film. Sedangkan untuk film yang kurang direkomendasikan, untuk ditonton berada di Cluster 1 dengan jumlah 1153 film. Dari hasil klaster tersebut dapat merekomendasikan film yang mungkin bisa menjadi referensi untuk ditonton.

Adapun saran untuk mengembangkan penelitian ini adalah melakukan pengklasteran dengan atribut lain selain *users_rating* dan *votes*. Penggunaan atribut lain bisa menghasilkan rekomendasi film berdasarkan preferensi yang diinginkan. Tak hanya itu, disarankan agar penelitian menggunakan data yang memiliki atribut yang bisa dilakukan pengklasteran serta tidak memiliki atribut *null* yang banyak.

5. DAFTAR RUJUKAN

[1] Indriani, U., 2018. Pendekatan K-Means Clustering Terhadap Rekomendasi Film untuk Ditonton. *Jurnal Sistem Informasi Kaputama (JSIK)*, 2 (1), pp. 84-88

[2] Pemerintah Kabupaten Pati, 2014. Sejarah Perkembangan Film Indonesia [Online] (Updated 10 Mar 2014) Available at: https://www.patikab.go.id/v2/id/2010/01/24/sejarah-perkembangan-film-indonesia/ [Accessed 8 Juni 2022]

[3] Dinata, R.K. et all, 2020. Analisis K-Means Clustering pada Data Sepeda Motor. *INFORMAL*, 5 (1), pp. 10-17

[4] Bastian, A., Sujadi, H., Febrianto, G., 2018. Penerapan Algoritma K-Means Clustering Analysis pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka). *Jurnal Sistem Informasi (Journal of Information System)*, 14 (1), pp. 26-32

[5] Priati and Fauzi A., 2017. *Data Mining dengan Teknik Clustering Menggunakan Algoritma K-Means pada Data Transaksi Superstore*. In: Seminar Nasional Informatika dan Aplikasinya (SNIA). Cimahi, 27 September 2017, Indonesia : Cimahi.

[6] Medium, 2017. Mempelajari Modeling Cross-Industry Standard Process for Data Mining atau CRISP-DM [Online] (Updated 17 Aug 2021) Available at: https://ruthsitorus.medium.com/mempelajari-modeling-cross-industry-standard-process-for-data-mining-atau-crisp-dm-166735c14159 [Accessed 22 Juni 2022]

[7] Rahmawati, L. et all, 2014. Analisa Clustering Menggunakan Metode K-Means dan Hierarchical Clustering (Studi Kasus : Dokumen Skripsi Jurusan Kimia, FMIPA, Universitas Sebelas Maret). *JURNAL ITSMART: Jurnal Teknologi dan Informasi*, 3(2), pp. 66-73

[8] Dewi, D.A.I.C., Pramita, D.A.K., 2019. Analisis Perbandingan Metode Elbow dan Sillhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali. *JURNAL MATRIX*, 9(3), pp. 102-109

[9] Anggara, M., Sujiani, H., Nasution, H., 2016. Pemilihan Distance Measure Pada K-Means Clustering Untuk Pengelompokan Member Di Alvaro Fitness. *Jurnal Sistem dan Teknologi Informasi (JUSTIN)*, 1 (1), pp. 1-6

[10] Paembonan, S., Abduh, H., 2021. Penerapan Metode Silhouette Coefficient Untuk Evaluasi Clustering Obat. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, 6 (1), pp. 48-54