

# Eksplorasi Pengolahan Citra *Thumbnail YouTube* Menggunakan Analisis *Clustering*

Elfira Rahma Putri<sup>a</sup>, David Aristoteles Kevas<sup>b</sup>, Greiva Viandra Zahrani<sup>c</sup>,  
Muhammad Farras Reswara Aryandra<sup>d</sup>

<sup>a</sup>(162012133010) Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya

<sup>b</sup>(162012133025) Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya

<sup>c</sup>(162012133053) Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya

<sup>d</sup>(162012133055) Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya

---

## Abstrak

YouTube menjadi salah satu platform video *online* yang sedang populer bagi para pelaku bisnis untuk menjangkau target pemasaran yang lebih luas dan menyeluruh. Kesadaran akan keuntungan hasil pemasaran melalui *platform* YouTube menyebabkan persaingan yang semakin ketat. Tiap kreator mencoba untuk menonjolkan ciri khas dari video yang dibuat, salah satunya melalui pembuatan *thumbnail*. *Thumbnail* sebagai kesan pertama bagi para pengguna dalam memilih video harus bersifat informatif dan menarik. Dalam penelitian ini akan dilakukan eksplorasi proses *image mining* dengan menggunakan analisis *clustering* dalam rangka mengidentifikasi karakteristik citra yang ada pada tiap *cluster*. Data yang digunakan dalam penelitian ini terdiri dari 334 *thumbnail* YouTube terkait konten pembelajaran bahasa pemrograman *Python* tingkat pemula yang diunggah dari tahun 2020 hingga 2022. Pertama, dilakukan analisis *clustering* secara langsung terhadap citra *thumbnail*. Kedua, dilakukan analisis *clustering* terhadap *visual attributes* dari citra *thumbnail*, meliputi skor *brightness*, *complexity*, *quality*, *colorfulness*, dan *detail views*. Dari analisis tersebut didapatkan hasil bahwa kebaikan model *clustering* pada data *visual attributes* dari citra *thumbnail* memiliki nilai kebaikan model *clustering* yang lebih baik dibandingkan dengan analisis *cluster* secara langsung terhadap citra *thumbnail*. Hasil terbaik didapatkan dengan menggunakan algoritma *Agglomerative* yang menghasilkan 2 *cluster*. Analisis karakteristik pada tiap *cluster* dilakukan dengan membandingkan hasil visualisasi *radar chart* berdasarkan ukuran standard deviasi, mean, dan maksimum dari *visual attributes* yang digunakan.

Kata Kunci : *Clustering, Thumbnail, Visual Attributes, YouTube*

---

## 1. Pendahuluan

Berdasarkan survei yang telah dilakukan oleh Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), tingkat penetrasi internet di Indonesia pada tahun 2021 telah mencapai angka 77.02% yang artinya 210.026.769 jiwa dari total populasi di Indonesia telah terkoneksi dengan internet (APJII, 2022). Hal tersebut secara jelas memberi gambaran terkait eratnya kehidupan masyarakat dengan teknologi media digital. *Youtube* sebagai sebuah situs web milik Google memungkinkan pengguna untuk mengunggah, menonton, dan berbagi video melalui *platform* ini. Per Mei 2019, YouTube menjadi salah satu platform video *online* terpopuler dengan total pengguna lebih dari 2 M per bulan (Spangler, 2019). YouTube telah menjadi media pemasaran populer bagi pelaku bisnis di berbagai bidang untuk menjangkau target pelanggan yang lebih luas, termasuk pada bidang pendidikan.

Berbagai konten edukasi pembelajaran telah diproduksi secara individual maupun dibawah naungan sebuah lembaga pendidikan. *Thumbnail* sebagai kesan pertama bagi para pengguna dalam mengakses video YouTube menjadi salah satu hal yang menarik untuk dianalisis. Saat menjelajahi YouTube, pengguna memutuskan video yang akan ditonton berdasarkan informasi dan kesan pertama yang ditangkap melalui *thumbnail* video tersebut. Menurut salah satu konten kreator YouTube yang telah berpartisipasi lebih dari satu dekade, Edho Zell menyatakan bahwa seorang kreator YouTube harus mengerti bagaimana sudut pandang dari penonton beserta dengan kebutuhannya (Afifyah, S., 2022). Berdasarkan

penelitian yang dilakukan oleh Koh, B., dkk (2022), menyatakan bahwa terdapat hubungan antara atribut visual *thumbnail* dengan pencapaian jumlah tayangan video. Sehingga, informasi yang dicantumkan pada *thumbnail* ataupun judul video harus diiringi dengan visual yang menarik dan dapat meningkatkan potensi pengguna dalam memilih video.

Oleh karena itu, penelitian ini dilakukan untuk menganalisis serangkaian proses pengolahan citra atau *image processing* dari data citra *thumbnail* YouTube. Metode analisis *clustering* digunakan untuk menganalisis bagaimana karakteristik citra *thumbnail* pada tiap *cluster* yang terbentuk. Analisis *clustering* akan mempartisi data ke dalam beberapa *cluster*, sehingga citra dengan karakteristik yang sama akan dikelompokkan ke dalam satu *cluster*, dan citra dengan karakteristik berbeda dikelompokkan ke dalam *cluster* lainnya. Dalam suatu *cluster*, antar data bersifat similar atau mirip satu sama lain, sedangkan bersifat dissimilar atau tidak mirip dengan data pada *cluster* lainnya. Oleh karena itu, dengan melakukan analisis *clustering* akan didapatkan informasi mengenai pola distribusi data secara keseluruhan ataupun menemukan hubungan keterkaitan yang menarik antar atribut data yang digunakan. Pada penelitian ini akan digunakan beberapa algoritma *clustering*, meliputi K-Means, BIRCH, *Agglomerative*, *Mean Shift*, dan OPTICS. Evaluasi kebaikan model *clustering* akan dianalisis melalui dua kriteria, yakni skor *Silhouette* dan skor *Davies Bouldin*.

## 2. Landasan Teori

### 2.1 Thumbnail

Menurut Oxford Learner's Dictionary, kata *thumbnail* aslinya berarti “kuku pada ibu jari”. Sekarang kata *thumbnail* secara luas diartikan sebagai “gambar yang sangat kecil di layar komputer yang menunjukkan seperti apa gambar yang lebih besar, atau seperti apa halaman dokumen ketika dicetak”. Pada aplikasi YouTube, *thumbnail video* merupakan gambar yang menampilkan sinopsis dari suatu video (Chu, 2020).

### 2.2 Citra Gambar

Menurut Merriam-Webster, citra gambar (*image*) merupakan sebuah representasi visual dari sebuah objek. Sebuah citra gambar dapat direpresentasikan oleh sebuah matriks dua dimensi dengan N baris dan M kolom. Perpotongan antara baris dan kolom tersebut merupakan elemen terkecil dari sebuah citra yang disebut sebagai piksel (*pixel*) (Hernando, dkk., 2020).

$$F = [f(i, j)] = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,M) \\ f(1,0) & f(1,1) & \dots & f(1,M) \\ \vdots & \vdots & \ddots & \vdots \\ f(N-1,0) & f(N-1,1) & \dots & f(N-1,M-1) \end{bmatrix}$$

Gambar 1. Representasi dari citra

### 2.3 Analisis Clustering

Analisis cluster merupakan analisis yang dilakukan untuk mengelompokkan data berdasarkan informasi-informasi dan hubungan yang ditemukan diantara data-datanya.

#### 2.3.1 Principal Component Analysis (PCA)

Menurut Jafarzadegan, dkk (2019), *Principal Component Analysis* atau analisis komponen utama merupakan suatu teknik mereduksi data dan memiliki tingkat kesalahan yang lebih rendah daripada metode reduksi dimensi lainnya.

#### 2.3.2 K-Means Clustering

Menurut Agusta (2007), *k-Means* merupakan salah satu metode clustering *nonhierarki* yang mengelompokkan data menjadi satu atau lebih *cluster*. Data yang memiliki karakteristik yang sama akan dikelompokkan dalam satu *cluster*, sedangkan yang karakteristiknya berbeda akan dikelompokkan dengan *cluster* yang lain (Ong, 2013).

#### 2.3.3 Mean Shift

Menurut Anand, dkk (2013), *Mean Shift* merupakan algoritma pengklasteran mode yang populer, yang menempatkan mode dalam catatan dengan memaksimalkan perkiraan kepadatan kernel (KDE)

secara iteratif. Berbeda dengan pendekatan clustering K-means klasik, tidak ada asumsi yang melekat pada bentuk distribusi atau jumlah mode/cluster (Derpanis, 2005).

#### 2.3.4 BIRCH

BIRCH berurusan dengan kumpulan data besar dengan menghasilkan ringkasan yang lebih padat terlebih dahulu dan mempertahankan informasi distribusi sebanyak mungkin, dan kemudian mengklasterkan ringkasan data, bukan data asli (Zhang, dkk., 1997).

#### 2.3.5 Agglomerative

Algoritma agglomerative clustering adalah proses pengelompokan secara *bottom-up*. Awalnya, setiap objek membentuk clusternya masing-masing. Pada setiap langkah selanjutnya, dua cluster terdekat akan digabungkan hingga hanya tersisa satu cluster. Proses clustering ini termasuk kedalam metode clustering *hierarki* (Ackerman, dkk., 2014).

#### 2.3.6 OPTICS

OPTICS: Ordering Points To Identify the Clustering Structure atau yang berarti mengurutkan poin untuk mengidentifikasi struktur cluster merupakan algoritma baru yang bertujuan untuk menganalisis cluster yang tidak menghasilkan pengelompokan kumpulan data secara eksplisit; tetapi sebaliknya membuat urutan tambahan dari data yang mewakili struktur pengelompokan berbasis kepadatannya. Pengurutan cluster ini berisi informasi yang setara dengan pengelompokan berbasis kepadatan yang sesuai dengan berbagai pengaturan parameter (Ankerst, dkk., 1999).

### 2.4 Evaluasi Kebaikan Model Clustering

#### 2.4.1 Silhouette Score

Fungsi *silhouette score* pada scikit-learn berfungsi untuk menghitung koefisien rata-rata siluet dari semua sampel. Koefisien siluet dihitung dengan memperhitungkan rata-rata jarak intra-cluster  $a$  dan rata-rata jarak terdekat-cluster  $b$  untuk setiap titik data. Koefisien siluet untuk sebuah sampel adalah  $(b - a)/\max(a, b)$  (Shahapure, dkk., 2020).

- Skor *silhouette* dengan nilai mendekati +1 berarti titik data berada di cluster yang benar.
- Skor *silhouette* dengan nilai mendekati 0 berarti titik data mungkin termasuk dalam cluster lain.
- Skor *silhouette* dengan nilai mendekati -1 artinya, titik data berada di cluster yang salah.

#### 2.4.1 Davies-Bouldin Index (DBI)

Menurut Kovács, dkk (2005), Indeks Davies–Bouldin mengukur rata-rata kesamaan antara setiap cluster dan yang paling mirip. Indeks ini didasarkan pada ukuran kesamaan cluster ( $R_{ij}$ ) yang basisnya adalah ukuran dispersi suatu cluster ( $s_i$ ) dan ukuran ketidaksamaan cluster ( $d_{ij}$ ). Menurut Halkidi, dkk (2002),  $R_{ij}$  dan DBI dapat didefinisikan dengan rumus sebagai berikut,

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \quad DBI = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i, \text{ dimana } R_i = \max_{j \neq i} (R_{ij}), i = 1 \dots n_c$$

Nilai DBI yang didapatkan dari setiap cluster kemudian dibandingkan hasilnya, dengan cluster yang memiliki nilai DBI terkecil atau mendekati 0 namun tidak negatif adalah cluster yang paling optimal (Ashari, dkk., 2022).

### 3. Sumber Data dan Metodologi

#### 3.1 Sumber Data

Jenis data yang digunakan merupakan data primer, yaitu sumber data yang didapatkan secara langsung oleh peneliti (Sugiyono, 2016). Untuk proses pengambilan data digunakan metode *scraping* dengan library Selenium pada Python. Data yang diambil berasal langsung dari Youtube dengan kata kunci “Python for beginner” berupa data gambar yang merupakan foto *thumbnail*, *link*, jumlah *viewers*, dan tanggal video di-*upload*. Hasil *scraping* dengan *filter* tahun *upload* pada rentang 2020-2022 didapatkan jumlah gambar *thumbnail* sebanyak 334 foto. Dari hasil *scraping* didapatkan tiga variabel, yaitu *link*, detail *views*, dan *upload date*. Setelah itu, menambahkan variabel label yang berisi tiga kategori berdasarkan jumlah *viewers*. Untuk memproses analisis regresi, dilakukan ekstraksi *visual attributes* dari gambar *thumbnail* sehingga didapatkan empat variabel baru, yaitu *brightness*, *complexity*, *colorfulness*, dan *quality*.

Table 1. Variabel Penelitian

Variabel	Keterangan
Link	Link Youtube <i>thumbnail</i>
Upload Date	Tanggal <i>upload</i> video
Detail Views	Jumlah <i>viewers</i> secara spesifik (numerik)
Brightness	Skor kecerahan gambar yang didapatkan dari rumus $B = \sqrt{0.241Rr^2 + 0.391Gr^2 + 0.068Bl^2}$
Complexity	Skor kompleksitas gambar yang didapat dari <i>Random Forest Classifier</i> untuk mengklasifikasikan tepi sebagai sederhana atau kompleks
Colorfulness	Skor keberagaman warna yang didapat dari rumus $C = \sqrt{\sigma^2_{rg} + \sigma^2_{yb}} + 0.3 \sqrt{\mu^2_{rg} + \mu^2_{yb}}$
Quality	Skor kualitas gambar yang didapat menggunakan algoritma BRISQUE. Semakin rendah skor BRISQUE maka semakin baik kualitas gambar tersebut
Label	Kategori <i>thumbnail</i> berdasarkan jumlah <i>viewers</i> (Rendah, Sedang, dan Tinggi)

### 3.2 Metodologi

Langkah awal yang dilakukan pada penelitian ini adalah melakukan proses pelabelan berdasarkan jumlah *viewers*. Label dibagi menjadi tiga kelompok yaitu untuk jumlah *viewers* dalam rentang 100-24999 diberi label Rendah, untuk rentang 25000-249999 diberi label Sedang, dan untuk jumlah *viewers* yang lebih dari sama dengan 250000 diberi label Tinggi. setelah itu, divisualisasikan untuk melihat sebaran warna untuk masing-masing label.

*Preprocessing data* pertama, melakukan proses normalisasi dengan membagi data dengan 255 (nilai maksimum intensitas piksel), lalu dilakukan standarisasi untuk persiapan proses PCA. PCA dilakukan untuk mereduksi data menjadi dua dimensi dengan ukuran 267 x 200. Hasil PCA ini akan digunakan analisis clustering pada data gambar.

Untuk analisis *clustering*, akan dibagi menjadi dua yaitu clustering data gambar dan clustering nilai *visual attributes*. Untuk clustering data gambar, digunakan beberapa algoritma, diantaranya algoritma K-Means dengan nilai k=8, algoritma BIRCH dengan nilai n cluster = 8, algoritma Mean Shift, algoritma agglomerative, dan algoritma OPTICS. Terakhir, dilakukan komparasi dengan nilai silhouette score dan nilai DBI untuk setiap algoritma.

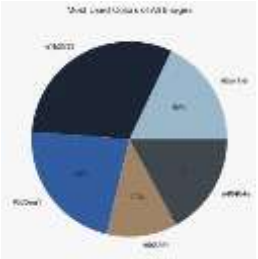
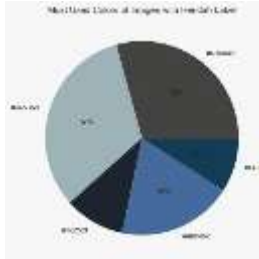
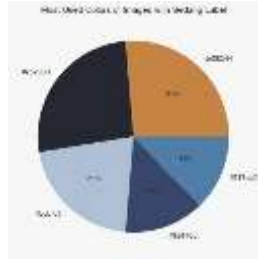
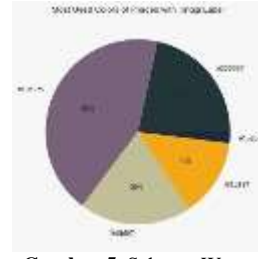
Langkah selanjutnya, untuk melakukan analisis clustering dengan nilai *visual attributes*. Maka, diperlukan menghitung nilai *brightness* yang merupakan skor kecerahan gambar yang didapatkan dari rumus  $B = \sqrt{0.241Rr^2 + 0.391Gr^2 + 0.068Bl^2}$ , *complexity* yang merupakan skor kompleksitas gambar yang didapat dari *Random Forest Classifier* untuk mengklasifikasikan tepi sebagai sederhana atau kompleks, *colorfulness* yang merupakan skor keberagaman warna yang didapat dari rumus  $C = \sqrt{\sigma^2_{rg} + \sigma^2_{yb}} + 0.3 \sqrt{\mu^2_{rg} + \mu^2_{yb}}$ , dan *quality* yang merupakan skor kualitas gambar yang didapat menggunakan algoritma BRISQUE. Lalu, variabel Detail Views, brightness, complexity, colorfulness, dan quality dianalisis clustering untuk mengetahui karakteristik citra thumbnail dari cluster yang terbentuk. Algoritma yang digunakan diantaranya, algoritma K-Means dengan nilai k=4, algoritma BIRCH dengan nilai n cluster = 4, algoritma Mean Shift, algoritma agglomerative, dan algoritma OPTICS. Lalu, dilakukan komparasi dengan nilai silhouette score dan nilai DBI untuk setiap algoritma. Selanjutnya, melakukan visualisasi *radar chart* untuk mengetahui karakteristik setiap *cluster* jika dilihat dari lima variabel tersebut dengan tiga fungsi statistika deskriptif, yaitu *standard deviation*, *maximum*, dan median. Terakhir, diambil kesimpulan dari penelitian.

## 4. Analisis dan Pembahasan

### 4.1 Eksplorasi dan Visualisasi Data

#### 4.1.1 Persentase Komponen Warna

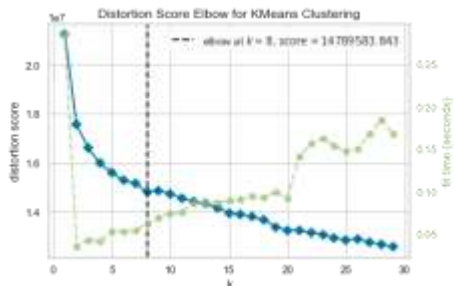
Table 2. Persentase Komponen Warna

Semua Data	Label Rendah	Label Sedang	Label Tinggi
 <p><b>Gambar 2.</b> Sebaran Warna Semua Data</p>	 <p><b>Gambar 3.</b> Sebaran Warna Data Berlabel Rendah</p>	 <p><b>Gambar 4.</b> Sebaran Warna Data Berlabel Sedang</p>	 <p><b>Gambar 5.</b> Sebaran Warna Data Berlabel Tinggi</p>
<p>Warna gelap (#1b2533) dengan nilai RGB sebesar (27,37,51) cenderung mendominasi komponen warna gambar pada semua gambar <i>thumbnail</i> sebesar 31%.</p>	<p>Warna #9eb3b8 dengan nilai RGB sebesar (158,179,184) cenderung mendominasi komponen warna gambar pada gambar <i>thumbnail</i> dengan label Rendah sebesar 32%.</p>	<p>Warna #252830 dengan nilai RGB sebesar (37,40,48) dan warna #c58344 dengan nilai RGB sebesar (64,126,201) cenderung mendominasi komponen warna gambar pada gambar <i>thumbnail</i> dengan label Sedang sebesar 26%.</p>	<p>Warna #786279 dengan nilai RGB sebesar (130,86,132) cenderung mendominasi komponen warna gambar pada gambar <i>thumbnail</i> dengan label Tinggi sebesar 43%.</p>

## 4.2 Analisis Clustering

### 4.2.1 Hasil Analisis Clustering untuk Data Gambar dari Setiap Algoritma

#### 4.2.1.1 Penentuan Nilai $k$



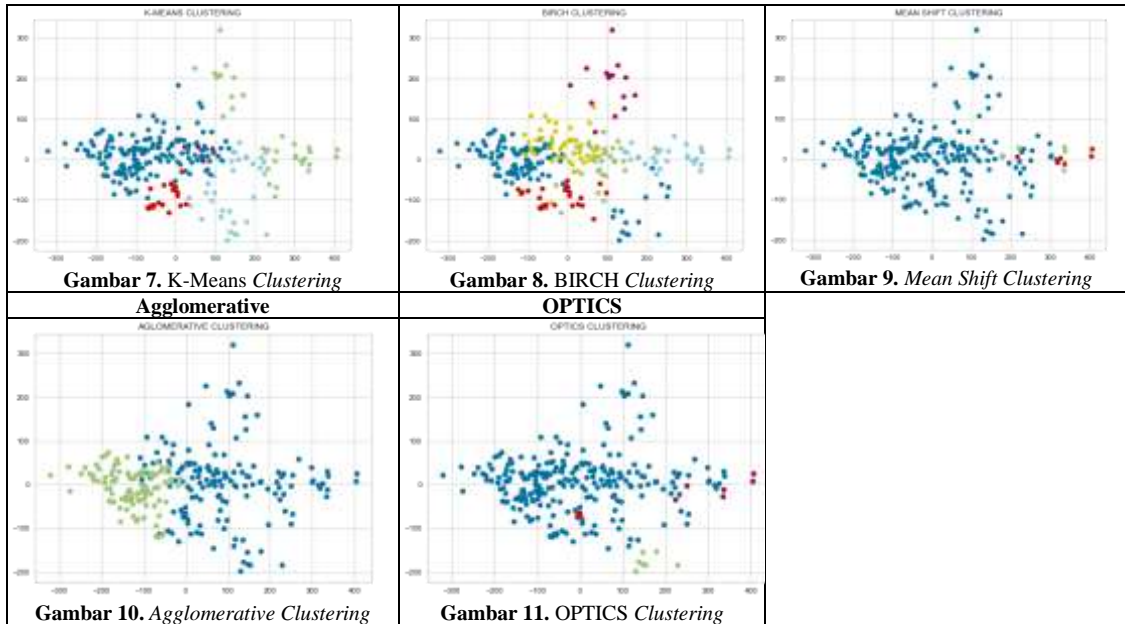
**Gambar 6.** Elbow Method

Untuk menentukan nilai  $k$  optimal yang akan digunakan dalam menentukan jumlah *cluster*, digunakan *elbow method*. Sesuai dengan visualisasi diatas, sumbu vertikal merupakan *distortion score* yang menyatakan ukuran titik data terhadap centroid kelompok, semakin kecil nilai *distortion score* berarti semakin baik pembagian kelompok tersebut. Didapatkan hasil bahwa jumlah  $k$  optimal adalah 8, yakni titik dimana terjadi penurunan yang signifikan.

#### 4.2.1.2 Hasil Visualisasi Setiap Algoritma

Table 3. Hasil Visualisasi Setiap Algoritma Data Gambar

Hasil Visualisasi Setiap Algoritma		
K-Means	BIRCH	Mean Shift



Hasil dari *cluster* menggunakan lima algoritma didapatkan pada algoritma K-Means terbentuk 8 *cluster*, 8 *cluster* pada algoritma BIRCH, 3 *cluster* pada algoritma *Mean Shift*, 2 *cluster* pada algoritma Agglomerative, dan 4 *cluster* pada algoritma OPTICS. Secara keseluruhan, setiap algoritma memiliki hasil clustering yang saling *overlap* atau tumpang tindih dilihat dari titik warna yang terlihat tercampur. Dengan demikian, dapat disimpulkan clustering yang dilakukan masih kurang baik.

#### 4.2.1.3 Hasil Nilai Matriks Evaluasi Setiap Algoritma

Table 4. Hasil Nilai Matriks Evaluasi

Algoritma	Silhouette Score	Nilai DBI
K-Means	0.0533	2.5456
BIRCH	0.0356	2.7541
Mean Shift	0.1669	1.9451
Agglomerative	0.1192	2.2853
OPTICS	-0.1288	1.4977

Silhouette score menunjukkan kinerja clustering pada data. Jika nilai semakin mendekati 1 maka pengelompokan objek dalam satu *cluster* semakin baik, sedangkan ketika nilai mendekati -1 maka pengelompokan objek dalam satu *cluster* semakin buruk dan ketika nilai mendekati nilai 0 maka sampel tersebut jelas tidak cocok. Dari tabel di atas, terlihat nilai silhouette tertinggi berada pada algoritma Mean Shift, yaitu sebesar 0.1669. Artinya, jika fokus utama penelitian pada pengelompokan objek dalam satu cluster, algoritma Mean Shift paling cocok digunakan.

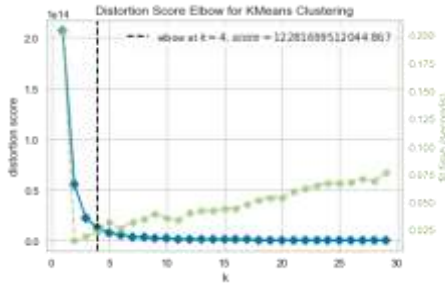
Nilai DBI mengukur seberapa terpisah kelompok-kelompok dalam data. Semakin rendah nilai DBI maka kinerja *cluster* semakin baik. Dari tabel di atas, terlihat bahwa nilai DBI terendah berada pada algoritma OPTICS, yaitu 1.4977. Artinya, jika fokus penelitian pada perbedaan kelompok-kelompok pada data maka algoritma OPTICS paling cocok digunakan.

Secara keseluruhan dapat dikatakan bahwa untuk clustering data gambar, algoritma Mean Shift merupakan algoritma yang paling cocok digunakan karena memiliki silhouette score terbesar dan nilai DBI

terkecil kedua dan hanya berbeda 0.447 saja. Cluster yang terbentuk dari algoritma tersebut sebanyak 2 *cluster*.

#### 4.2.2 Hasil Analisis Clustering untuk Data *Visual Attributes*

##### 4.2.2.1 Penentuan Nilai $k$

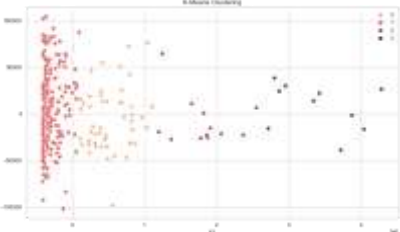
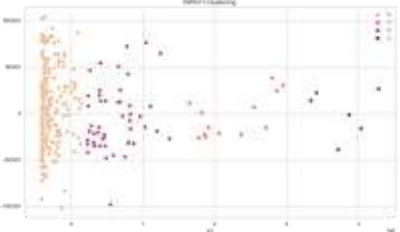
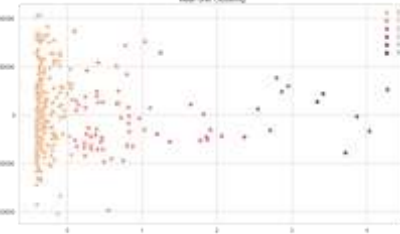
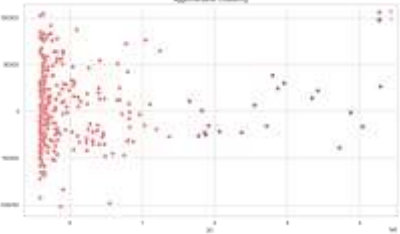


Gambar 12. Elbow Method

Untuk menentukan nilai  $k$  optimal yang akan digunakan dalam menentukan jumlah *cluster*, digunakan *elbow method*. Sesuai dengan visualisasi diatas, sumbu vertikal merupakan *distortion score* yang menyatakan ukuran titik data terhadap centroid kelompok, semakin kecil nilai *distortion score* berarti semakin baik pembagian kelompok tersebut. Didapatkan hasil bahwa jumlah  $k$  optimal adalah 4, yakni titik dimana terjadi penurunan yang signifikan.

##### 4.2.2.2 Hasil Visualisasi Setiap Algoritma

Table 5. Hasil Visualisasi Setiap Algoritma Data *Visual Attributes*

Hasil Visualisasi Setiap Algoritma	
K-Means	BIRCH
 <p><b>Gambar 13. K-Means Clustering</b></p> <p>Dengan menggunakan <math>k=4</math>, didapatkan hasil yang cukup baik, yakni tiap titik data terletak sesuai dengan <i>cluster</i> nya, dan antar <i>cluster</i> cenderung heterogen. Terlihat bahwa mayoritas data berada pada <i>cluster</i> 1</p>	 <p><b>Gambar 14. BIRCH Clustering</b></p> <p>Dengan menggunakan <math>k=4</math>, didapatkan hasil yang cukup baik, yakni tiap titik data terletak sesuai dengan <i>cluster</i> nya, dan antar <i>cluster</i> cenderung heterogen. Terlihat bahwa mayoritas data berada pada <i>cluster</i> 0</p>
Mean Shift	Agglomerative
 <p><b>Gambar 15. Mean Shift Clustering</b></p> <p>Terbentuk 6 <i>cluster</i> dengan hasil yang cukup baik, yakni tiap titik data terletak sesuai dengan <i>cluster</i> nya, dan antar <i>cluster</i> cenderung heterogen. Terlihat bahwa mayoritas data berada pada <i>cluster</i> 0, dan data pada <i>cluster</i> 3,4,5 cenderung memiliki karakteristik yang sangat berbeda dengan <i>cluster</i> 1,2</p>	 <p><b>Gambar 16. Agglomerative Clustering</b></p> <p>Dengan proses <i>agglomerative (bottom-up)</i>, terbentuk 2 buah <i>cluster</i>. Visualisasi hasil <i>clustering</i> menggambarkan <i>cluster</i> yang terbentuk sudah baik, yakni tiap titik data terletak sesuai dengan <i>cluster</i> nya, dan antar <i>cluster</i> cenderung heterogen. Terlihat bahwa mayoritas data berada pada <i>cluster</i> 0</p>
OPTICS	



**Gambar 17.** OPTICS Clustering

Dengan menggunakan  $min\_samples=3$ , yakni jumlah data minimum yang diperlukan untuk menjadi *core point*, terbentuk 6 *cluster*. Dari visualisasi diatas, didapatkan hasil *clustering* yang kurang baik dikarenakan data antar *cluster* cenderung homogen.

#### 4.2.2.3 Hasil Nilai Matriks Evaluasi Setiap Algoritma

Table 6. Hasil Nilai Matriks Evaluasi

Algoritma	Silhouette Score	Nilai DBI
K-Means	0.7610	0.4472
BIRCH	0.7722	0.4378
Mean Shift	0.7587	0.4522
Agglomerative	0.8487	0.3388
OPTICS	0.1454	2.3236

Silhouette score menunjukkan kinerja clustering pada data. Jika nilai semakin mendekati 1 maka pengelompokan objek dalam satu *cluster* semakin baik, sedangkan ketika nilai mendekati -1 maka pengelompokan objek dalam satu *cluster* semakin buruk dan ketika nilai mendekati nilai 0 maka sampel tersebut jelas tidak cocok. Dari tabel di atas, terlihat nilai silhouette tertinggi berada pada algoritma Agglomerative, yaitu sebesar 0.8487. Artinya, jika fokus utama penelitian pada pengelompokkan objek dalam satu cluster, algoritma Mean Shift paling cocok digunakan.

Nilai DBI mengukur seberapa terpisah kelompok-kelompok dalam data. Semakin rendah nilai DBI maka kinerja *cluster* semakin baik. Dari tabel di atas, terlihat bahwa nilai DBI terendah berada pada algoritma OPTICS, yaitu 0.3388. Artinya, jika fokus penelitian pada perbedaan kelompok-kelompok pada data maka algoritma OPTICS paling cocok digunakan.

Secara keseluruhan dapat dikatakan bahwa untuk clustering data gambar, algoritma Agglomerative merupakan algoritma yang paling cocok digunakan karena memiliki silhouette score terbesar dan nilai DBI terkecil.

#### 4.2.2.4 Hasil Visualisasi Radar Chart Setiap Cluster

Table 7. Hasil Visualisasi Radar Chart

Analisis Karakteristik Cluster Berdasarkan		
Standar Deviasi	Maksimum	Median
<p><b>Gambar 18.</b> Dengan Standard Deviation</p> <p>Dapat disimpulkan, cluster 0 memiliki rentang nilai yang beragam pada variabel Detail Views, sedangkan pada cluster 1, rentangnya beragam pada variabel brightness</p>	<p><b>Gambar 19.</b> Dengan Maximum</p> <p>Dapat disimpulkan, nilai maksimum antara cluster 0 dengan cluster 1, cukup jauh. Perbedaan yang lain adalah nilai maksimal setiap variabel kecuali Detail Views lebih tinggi daripada cluster 0</p>	<p><b>Gambar 20.</b> Dengan Median</p> <p>Dapat disimpulkan bahwa perbedaan karakter tiap clusternya sangat ditentukan oleh variabel Detail Views</p>



## 5. Kesimpulan dan Saran

Berdasarkan metode-metode yang telah dilakukan pada data primer *thumbnail* youtube dengan kata kunci “*python for beginner*” dalam rentang 2020 hingga 2022, didapatkan kesimpulan sebagai berikut:

1. Warna dengan nuansa gelap adalah warna dominan yang digunakan pada keseluruhan data. Ada kecenderungan penggunaan warna terang pada thumbnail dengan label tinggi. Semakin rendah jumlah penonton video tersebut, warna yang digunakan juga semakin gelap.
2. Berdasarkan visualisasi yang telah dilakukan, dapat diambil kesimpulan tidak ada korelasi yang kuat antara jumlah penonton dengan variabel *brightness*, *complexity*, *colorfulness*, dan *quality*. Selain itu, masing-masing variabel memiliki sebaran data yang berbeda-beda.
3. Berdasarkan Elbow Method, data dapat dikelompokkan menjadi 3 kluster. Kluster yang terbentuk dibedakan berdasarkan kemiripan warnanya. Kluster pertama berisi thumbnail berwarna-warni, kluster kedua berisi thumbnail dengan warna terang, dan kluster ketiga berisi thumbnail dengan warna gelap. Kluster ketiga memiliki anggota paling banyak, artinya, thumbnail dengan kata kunci “*python for beginner*” mayoritas berwarna gelap.
4. Clustering pada data gambar masih belum menghasilkan cluster yang baik meskipun telah menggunakan beberapa macam algoritma. Hal ini kemungkinan besar disebabkan karena tidak ada perbedaan yang cukup mencolok pada data gambar.
5. Clustering pada data *Visual Attributes* memberikan hasil yang lebih baik ketimbang data gambar. Setiap algoritma yang dicoba menghasilkan nilai silhouette dan DBI yang cukup baik. Algoritma terbaik dalam melakukan cluster ada Agglomerative Clustering dengan jumlah clusternya adalah dua.
6. Dari hasil clustering yang telah dibuat, jika dilihat berdasarkan nilai standar deviasi, Detail Views dari cluster 0 lebih bervariasi, sementara pada cluster 1, yang bervariasi. Jika dilihat dari nilai maksimalnya, perbedaan Detail Views pada cluster 0 dan 1 terpaut cukup jauh. Terakhir, jika dilihat dari nilai median, semua variabel pada cluster 0 dan 1 memiliki nilai yang sangat serupa, kecuali pada variabel Detail Views. Dapat disimpulkan bahwa karakteristik yang membedakan antara cluster 0 dan cluster 1 adalah variabel Detail Views.

Dari penelitian yang telah dilakukan, didapatkan beberapa saran sebagai berikut,

1. Bagi para kreator YouTube, atribut visual *thumbnail* bukan menjadi satu-satunya faktor yang mempengaruhi tingkat tayangan dari video yang diunggah. Berdasarkan analisis *clustering* juga didapatkan kesimpulan bahwa pemilihan warna tidak mempengaruhi jumlah tayangan video. Selain itu, berdasarkan analisis skor BRISQUE pada tiap label klasifikasi, didapatkan informasi bahwa kualitas gambar mempengaruhi jumlah tayangan video yang diunggah.
2. Untuk penelitian selanjutnya, dapat dicoba *hyperparameter tuning* pada algoritma *clustering* yang telah dicoba sebelumnya, terutama DBSCAN yang hanya mampu membentuk satu cluster. Dengan demikian, diharapkan dapat meningkatkan nilai metrics yang digunakan, baik silhouette *score* maupun DBI.

## Daftar Pustaka

- Ackermann, M.R., Blömer, J., Kuntze, D. and Sohler, C., 2014. Analysis of agglomerative clustering. *Algorithmica*, 69(1), pp.184-215.
- Agusta, Y., 2007. K-means–penerapan, permasalahan dan metode terkait. *Jurnal Sistem dan informatika*, 3(1), pp.47-60.
- Anand, S., Mittal, S., Tuzel, O. and Meer, P., 2013. Semi-supervised kernels mean shift clustering. *IEEE transactions on pattern analysis and machine intelligence*, 36(6), pp.1201-1215.
- Ankerst, M., Breunig, M.M., Kriegel, H.P. and Sander, J., 1999. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2), pp.49-60.
- Ashari, I.F., Banjarnahor, R., Farida, D.R., Aisyah, S.P., Dewi, A.P. and Humaya, N., 2022. Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies. *Journal of Applied Informatics and Computing*, 6(1), pp.07-15.

- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 2017. *Classification and regression trees*. Routledge.
- Chu, W.T., 2020. Session details: Attractiveness Computing in Multimedia. In *MMArt&ACM@ ICMR*.
- Cristianini, N. and Shawe-Taylor, J., 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press..
- Derpanis, K.G., 2005. Mean shift clustering. *Lecture Notes*, 32.
- Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2002. Clustering validity checking methods: Part II. *ACM Sigmod Record*, 31(3), pp.19-27.
- Hernando, D., Widodo, A.W. and Dewi, C., 2020. Pemanfaatan Fitur Warna dan Fitur Tekstur untuk Klasifikasi Jenis Penggunaan Lahan pada Citra Drone. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer E-ISSN*, 2548(2).
- Jafarzadegan, M., Safi-Esfahani, F. and Beheshti, Z., 2019. Combining hierarchical clustering approaches using the PCA method. *Expert Systems with Applications*, 137, pp.1-10.
- Koh, B. and Cui, F., 2022. An Exploration of the Relation between the Visual Attributes of Thumbnails and the View-Through of Videos: The Case of Branded Video Content.
- Kovács, F., Legány, C. and Babos, A., 2005, November. Cluster validity measurement techniques. In *6th International symposium of hungarian researchers on computational intelligence* (Vol. 35).
- Ong, J.O., 2013. Implementasi algoritma k-means clustering untuk menentukan strategi marketing president university.
- Shahapure, K.R. and Nicholas, C., 2020, October. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 747-748). IEEE.
- Spangler, T. 2019. YouTube now has 2 billion monthly users, who watch 250 million hours on TV screens daily. The Star Online, May 6, 2019. Retrieved from: <https://www.thestar.com.my/tech/tech-news/2019/05/06/youtube-now-has-2-billionmonthly-users-who-watch-250-million-hours-on-tv-screens-dail>
- Zhang, T., Ramakrishnan, R. and Livny, M., 1997. BIRCH: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1(2), pp.141-182.