“Perancangan Model Estimasi Viewers Youtube untuk Meningkatkan Popularitas Channel Prodi MR ITB”

Diva Awanisa Nahdi
14418033

01

BUSINESS UNDERSTANDING

Latar Belakang

Prodi MR ingin **meningkatkan popularitas** channel Youtube

Salah satu indikator popularitas: **jumlah view**

Akan dicari **faktor-faktor yang memengaruhi** jumlah view menggunakan model dari *task estimation*

Rumusan Masalah

“Bagaimana model estimasi *view* Youtube yang dapat meningkatkan popularitas channel prodi MR ITB?”

1. Bagaimana karakteristik dari data yang digunakan?
2. Bagaimana variabel-variabel yang paling berpengaruh ditentukan?
3. Bagaimana cara memilih model estimasi terbaik?

Asumsi

1. Data berdistribusi normal
2. Data memenuhi pola homoskedasitas
3. Data berpengaruh secara linear

Batasan

1. **Data yang digunakan** adalah data *Trending* Youtube Indonesia dari Bulan Februari sampai awal Desember 2021
2. Pengolahan data dilakukan menggunakan **bahasa Python**

02

DATA UNDERSTANDING

Data yang Digunakan

Data Trending Youtube Indonesia dengan 54.029 baris dan 27 kolom

video_id	object
publish_time	object
channel_id	object
title	object
description	object
thumbnail_url	object
thumbnail_width	float64
thumbnail_height	float64
channel_name	object
tags	object
category_id	int64
live_status	object
local_title	object
local_description	object
duration	object
dimension	object
definition	object
caption	bool
license_status	bool
allowed_region	object
blocked_region	object
view	float64
like	float64
dislike	float64
favorite	int64
comment	float64
trending_time	object

Data diekstrak menggunakan *unicode* UTF-8 dan *delimiter* berupa koma (,) agar dapat diolah dalam IDE Python

- Object: Data bertipe karakter
- Float: Data numerik berupa bilangan asli
- Integer: Data numerik berupa bilangan bulat
- Bool : Data logika *True* atau *False*

Ringkasan Statistik

Untuk data bertipe numerik

	thumbnail_width	thumbnail_height	category_id	view	like	dislike	favorite	comment
count	937.0	937.0	54059.000000	5.405100e+04	5.343000e+04	53430.000000	54059.0	5.380400e+04
mean	480.0	360.0	18.551287	3.675736e+06	1.812868e+05	4557.326932	0.0	1.813808e+04
std	0.0	0.0	7.094194	1.246979e+07	6.940099e+05	17018.392696	0.0	1.707775e+05
min	480.0	360.0	1.000000	1.528200e+04	1.600000e+01	0.000000	0.0	0.000000e+00
25%	480.0	360.0	10.000000	3.430485e+05	7.923500e+03	190.000000	0.0	5.790000e+02
50%	480.0	360.0	22.000000	9.263850e+05	2.478350e+04	577.000000	0.0	1.910500e+03
75%	480.0	360.0	24.000000	2.421131e+06	9.053700e+04	2001.000000	0.0	6.230000e+03
max	480.0	360.0	29.000000	3.359576e+08	1.700526e+07	433282.000000	0.0	7.195733e+06

Kolom hanya memiliki satu *value*

Kolom tidak memiliki nilai

Eksplorasi Data

Data diubah terlebih dahulu agar eksplorasi data dapat dilakukan dengan lebih baik

1. Pengubahan tipe data menjadi ‘datetime’ untuk kolom **publish_time** dan **trending_time**
2. Pembuatan kolom baru berupa **publish_day** (hari video diunggah), **publish_day_no** (hari video diunggah dalam bentuk numerik), **publish_hour** (waktu video diunggah), dan **trend_time** (tanggal trending)
3. Pembuatan kolom **jarak_trending** yang merupakan umur video saat trending dalam hari
4. Pembuatan kolom **description_length** = panjang tulisan pada kolom deskripsi
5. Pembuatan kolom **title_length** = panjang tulisan judul
6. Pembuatan kolom **no_of_tags** = jumlah tags yang dilampirkan
7. Pembuatan kolom **capitalized_word** = apakah judul memiliki huruf kapital
8. Pengubahan kolom **duration** dari bentuk ISO menjadi total detik

Eksplorasi Data

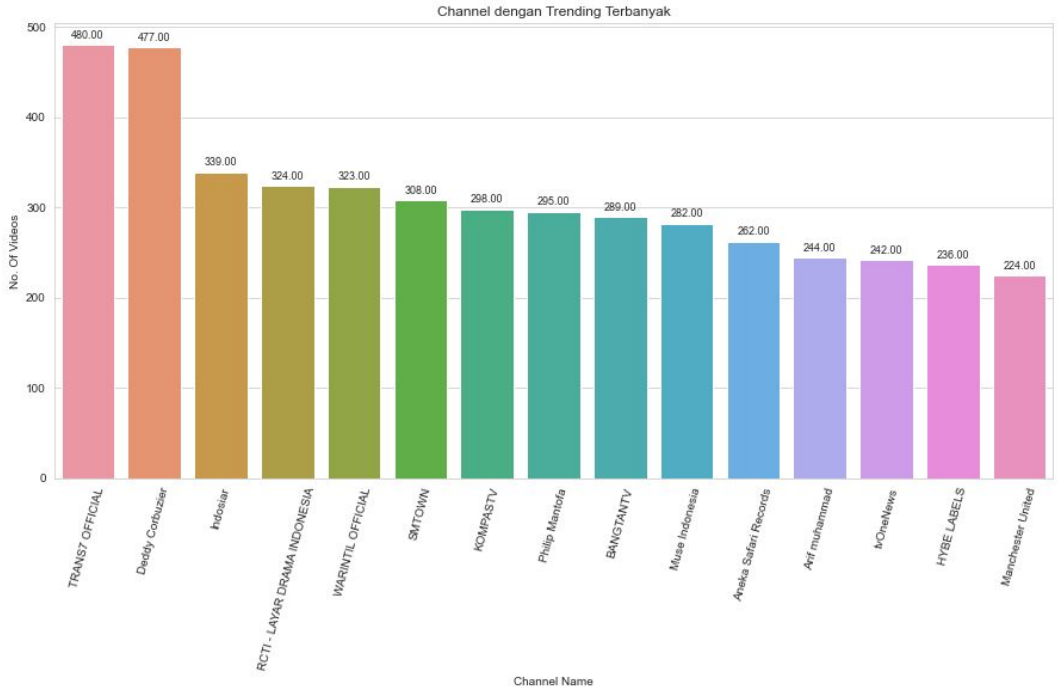
Data diubah terlebih dahulu agar eksplorasi data dapat dilakukan dengan lebih baik

Pengubahan nama kategori

ID	Category name
1	Film & Animation
2	Autos & Vehicles
10	Music
15	Pets & Animals
17	Sports
19	Travel & Events
20	Gaming
22	People & Blogs
23	Comedy
24	Entertainment
25	News & Politics
26	Howto & Style
27	Education
28	Science & Technology
29	Nonprofits & Activism

Eksplorasi Data

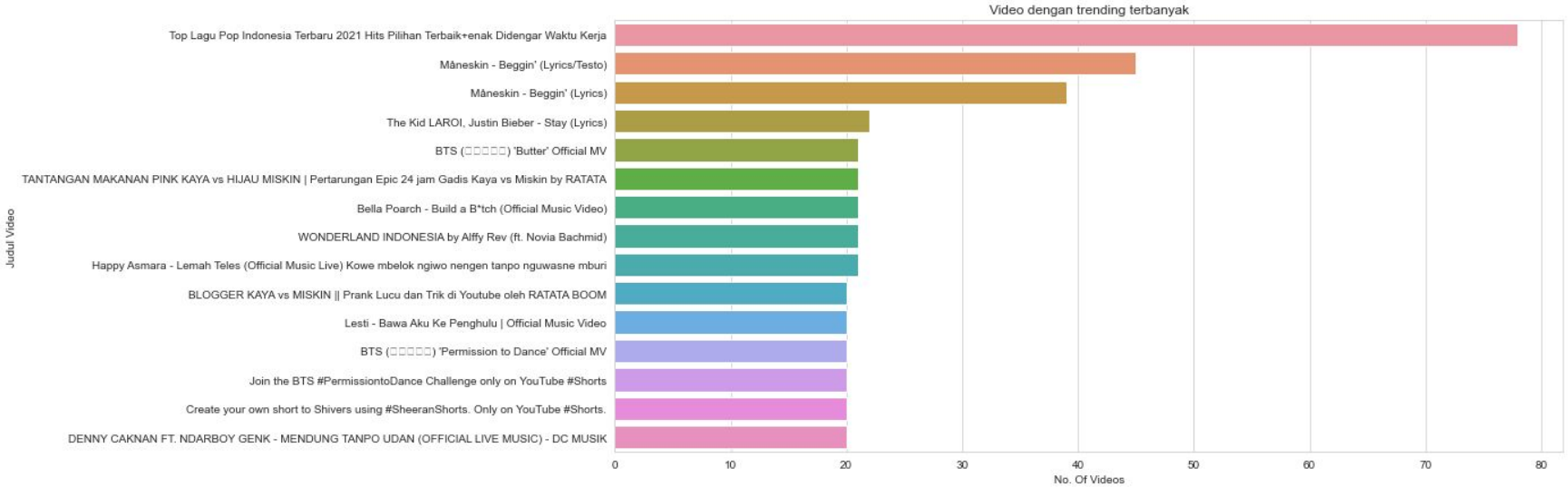
Channel dengan waktu trending terbanyak



Top 15 channel dengan waktu trending terbanyak. Banyak channel ini meliputi **channel dari TV stasiun lokal** di Indonesia

Eksplorasi Data

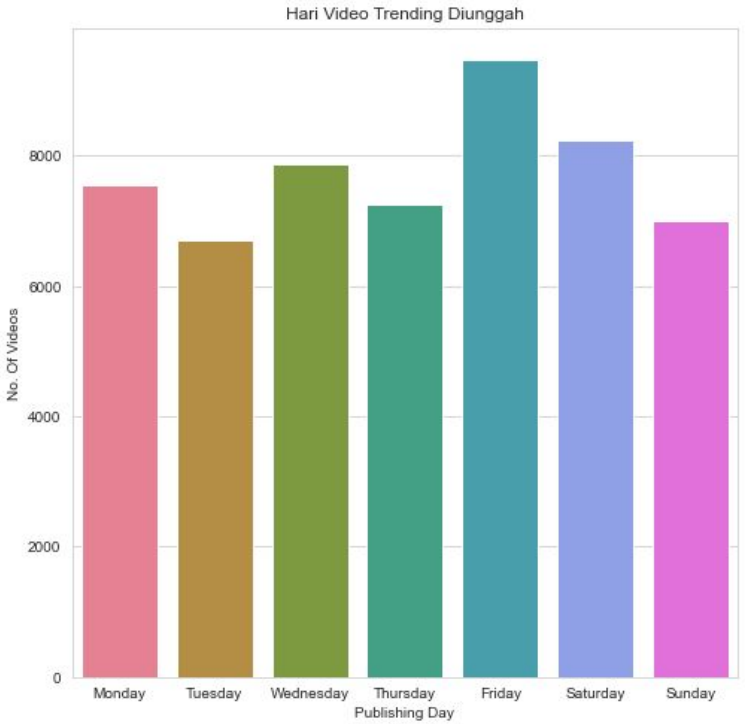
Video dengan waktu trending terbanyak



- Lima video yang paling sering *trending* di Indonesia adalah **video musik**
- Terdapat video yang *trending* sampai hampir 80 hari

Eksplorasi Data

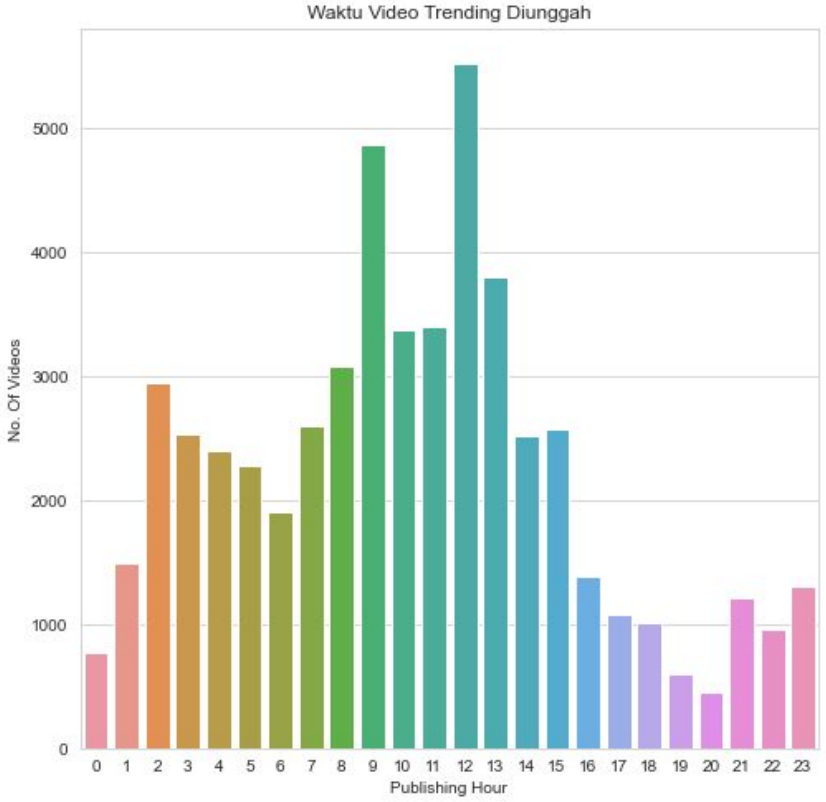
Hari video diunggah



- **Tidak ada distribusi** tertentu yang terlihat secara signifikan
- Video paling banyak diunggah pada hari Jumat dan paling sedikit pada hari Selasa

Eksplorasi Data

Waktu video diunggah



Video diunggah pada **masa aktif atau masa kerja** pengguna

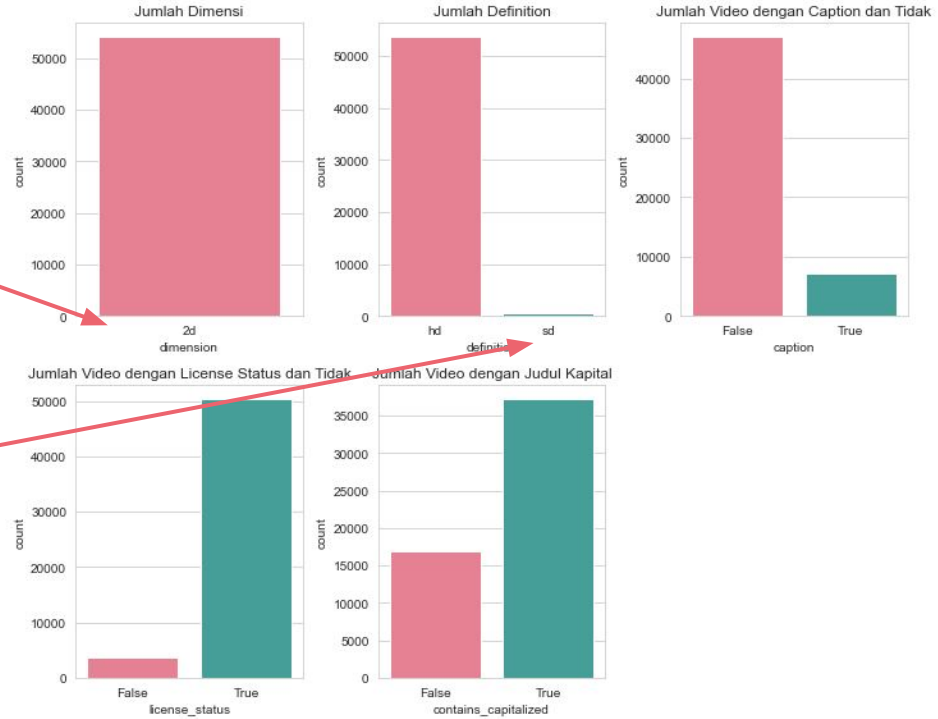
- Paling tinggi di rentang 9.00-13.00
- Paling rendah saat waktu istirahat, yaitu di rentang 16.00-23.00 dan 1.00-6.00

Eksplorasi Data

Jumlah setiap variabel kategori

Kolom hanya memiliki satu *value*

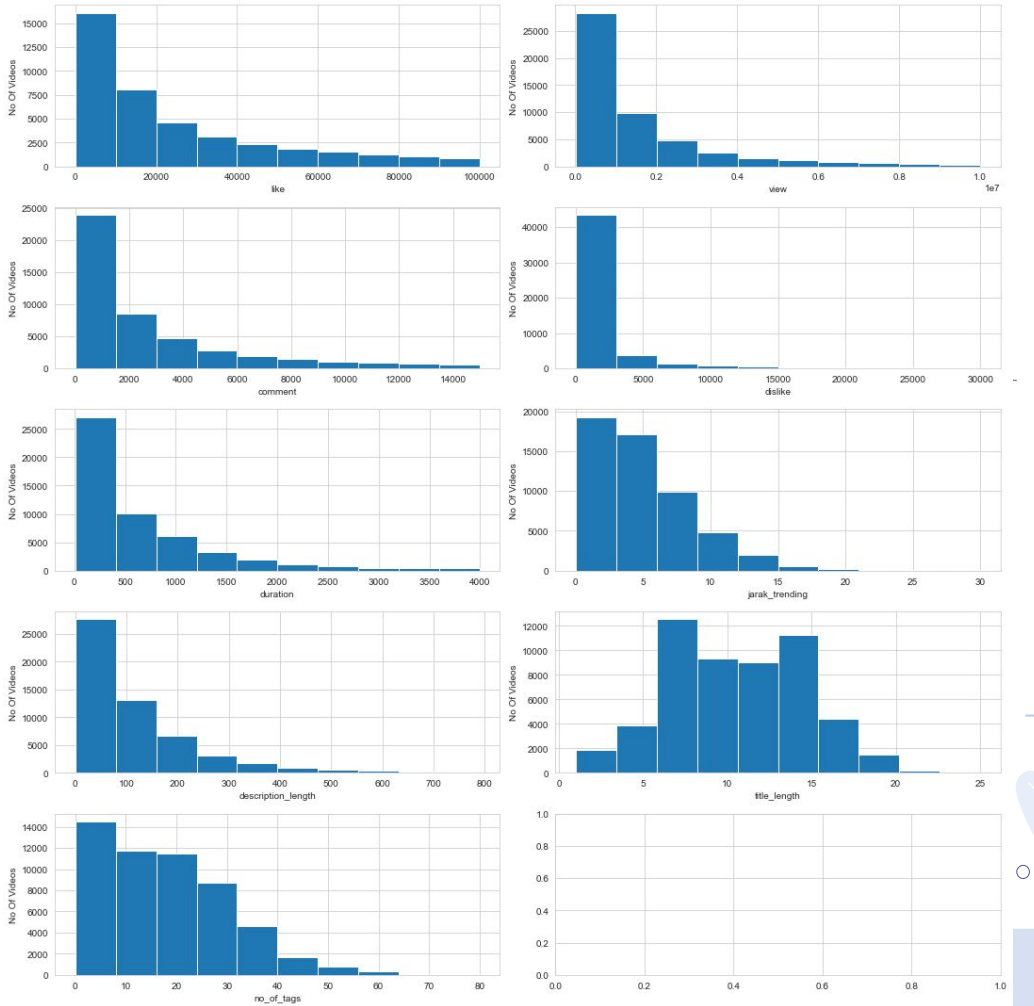
Nilai Sd sangat kecil



Eksplorasi Data

Distribusi variabel numerik

Semua variabel *Skewed* ke kanan kecuali variabel title_length



Eksplorasi Data

Kata paling banyak pada kolom deskripsi

Eksplorasi Data

Kata paling banyak pada kolom judul

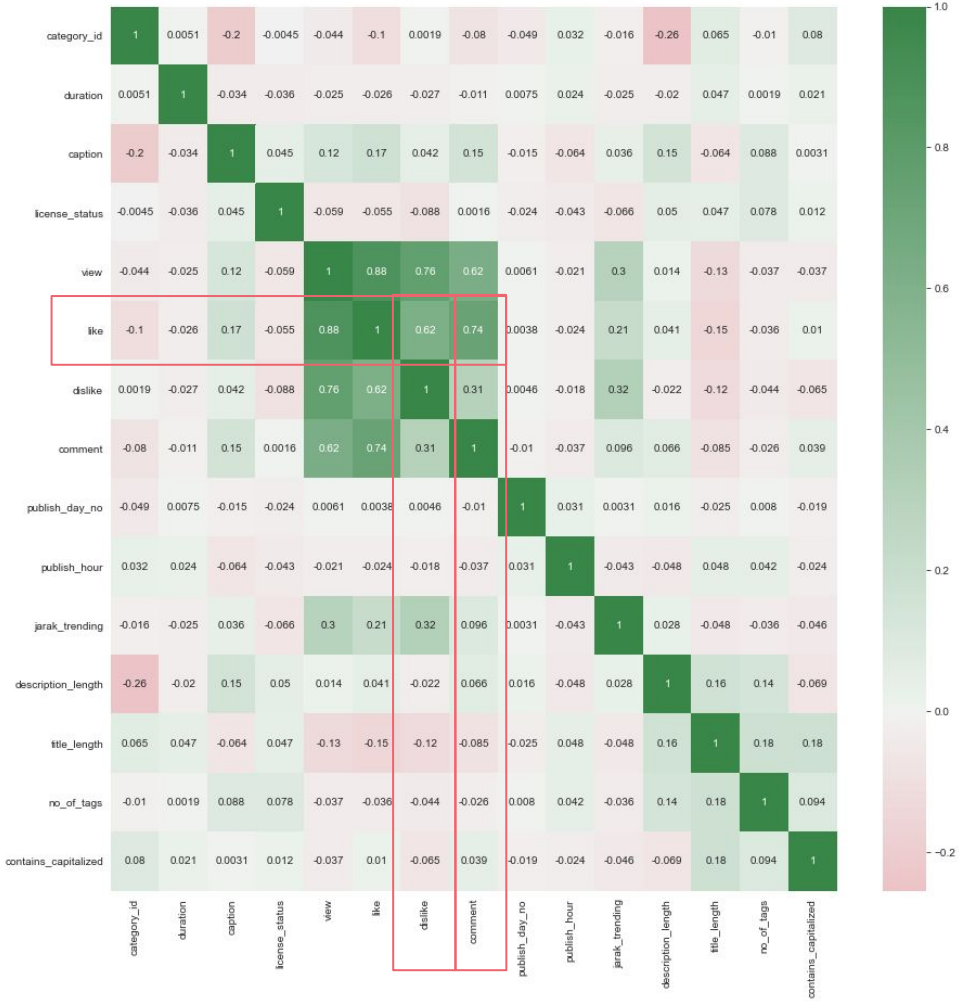
Dua kata paling banyak pada kolom judul

Video yang *trending* kebanyakan memiliki judul video musik resmi

Eksplorasi Data

Korelasi Variabel Numerik

Like, dislike, dan comment memiliki korelasi yang sangat tinggi, sehingga berpotensi menghasilkan multikolinearitas



03

DATA PREPARATION

Pembersihan dan Pemilihan Data

Pengecekan Data Terduplikasi

```
1 print(data.duplicated().sum())
```

```
0
```

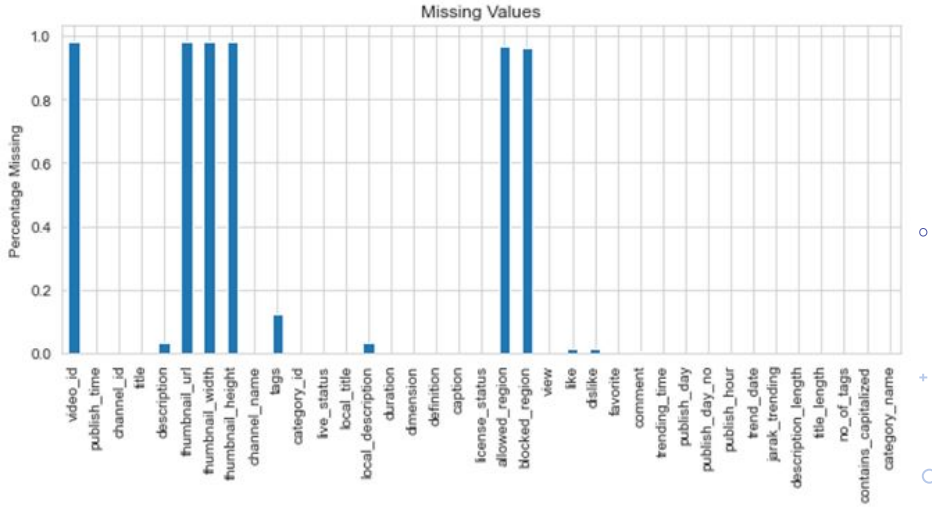
Tidak ada data terduplikasi

Pembersihan dan Pemilihan Data

Pengecekan Data yang Hilang

video_id	53122
publish_time	0
channel_id	0
title	0
description	1902
thumbnail_url	53122
thumbnail_width	53122
thumbnail_height	53122
channel_name	0
tags	6659
category_id	0
live_status	0
local_title	0
local_description	1902
duration	0
dimension	0
definition	0
caption	0
license_status	0
allowed_region	52248
blocked_region	51942

view	8
like	629
dislike	629
favorite	0
comment	255
trending_time	0
publish_day	0
publish_day_no	0
publish_hour	0
trend_date	0
jarak_trending	0
description_length	0
title_length	0
no_of_tags	0
contains_capitalized	0
category_name	0



kolom video_id, thumbnail_url, thumbnail_width, thumbnail_height, thumbnail_height, allowed_region, dan blocked_region **perlu dihapus** karena memiliki nilai yang hilang sangat besar

Pembersihan dan Pemilihan Data

Penghapusan Data

No	Kolom yang Dihapus	Alasan
1	Video_id	Terlalu banyak *missing data*
2	Publish_time	Sudah diwakili dengan kolom lain
3	Channel_id	Akan menghasilkan *dummies* yang terlalu banyak, menjadi batasan untuk tidak digunakan
4	Thumbnail_width	Hanya memiliki satu nilai
5	Thumbnail_height	Hanya memiliki satu nilai
6	Title	Akan menghasilkan *dummies* yang terlalu banyak, menjadi batasan untuk tidak digunakan. Diwakili oleh kolom panjang judul
7	Favorite	Hanya memiliki satu nilai
8	Thumbnail_url	Terlalu banyak data yang hilang
9	Tags	Sudah diwakili oleh no_of_tags
10	Description	Sudah diwakili oleh panjang deskripsi

Pembersihan dan Pemilihan Data

Penghapusan Data

No	Kolom yang Dihapus	Alasan
11	Channel_name	Akan menghasilkan *dummies* yang terlalu banyak, menjadi batasan untuk tidak digunakan
12	Live_status	Hanya memiliki satu nilai
13	Local_title	Akan menghasilkan *dummies* yang terlalu banyak, menjadi batasan untuk tidak digunakan. Diwakili oleh kolom panjang judul
14	Allowed_region	Terlalu banyak data yang hilang
15	Blocked_region	Terlalu banyak data yang hilang
16	Trend_date	Sudah diwakili oleh umur video
17	Dimension	Hanya memiliki satu nilai
18	Dislike	Berkorelasi tinggi dengan variabel independen lainnya
19	Comment	Berkorelasi tinggi dengan variabel independen lainnya
20	Definition	Memiliki nilai sd yang terlalu kecil, sehingga tidak akan menghasilkan pengaruh yang signifikan

Pembersihan dan Pemilihan Data

Penghapusan Data yang Hilang

duration	0
caption	0
license_status	0
view	8
like	629
publish_day_no	0
publish_hour	0
jarak_trending	0
description_length	0
title_length	0
no_of_tags	0
contains_capitalized	0
category_name	0

Baris dengan data yang hilang dihapus karena persentase nilai yang hilang **berada di bawah 5%**

Encoding Data

Data kategorkal nominal di *encode* secara *one-hot* encoding

duration	float64
caption	int64
license_status	int64
view	float64
like	float64
publish_day_no	int64
publish_hour	int64
jarak_trending	int64
description_length	int32
title_length	int64
no_of_tags	int32
contains_capitalized	int64
category_name_Comedy	uint8
category_name_Education	uint8
category_name_Entertainment	uint8
category_name_Film and Animation	uint8
category_name_Gaming	uint8
category_name_How to and Style	uint8
category_name_Music	uint8
category_name_News and Politics	uint8
category_name_Non Profits and Activism	uint8
category_name_People and Blogs	uint8
category_name_Pets and Animals	uint8
category_name_Science and Technology	uint8
category_name_Sport	uint8
category_name_Travel and Events	uint8

Category_name dijadikan variabel dummy untuk setiap kategorinya

Scaling Data

Data numerikal dinormalisasi agar analisis koefisien model dapat dilakukan dengan lebih mudah

	duration	caption	license_status	view	like	publish_day_no	publish_hour	jarak_trending	description_length	title_ler
count	53430.000000	53430.000000	53430.000000	53430.000000	53430.000000	53430.000000	53430.000000	53430.000000	53430.000000	53430.000
mean	0.001020	0.129964	0.934494	0.010929	0.010660	0.509879	0.442447	0.134632	0.126461	0.39
std	0.006874	0.336267	0.247419	0.037243	0.040812	0.326608	0.240285	0.111774	0.127056	0.164
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	0.000149	0.000000	1.000000	0.000976	0.000465	0.166667	0.260870	0.058824	0.042484	0.291
50%	0.000298	0.000000	1.000000	0.002715	0.001456	0.500000	0.434783	0.117647	0.082789	0.375
75%	0.000730	0.000000	1.000000	0.007186	0.005323	0.833333	0.565217	0.205882	0.171024	0.500
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000

Nilai maksimal menjadi 1 dan nilai minimal menjadi 0 untuk setiap kolom

Splitting Data

Data dipisah menjadi data train sebagai data untuk membangun model dan data test untuk mengevaluasi model

Ukuran *test size* adalah 20% dari dataset

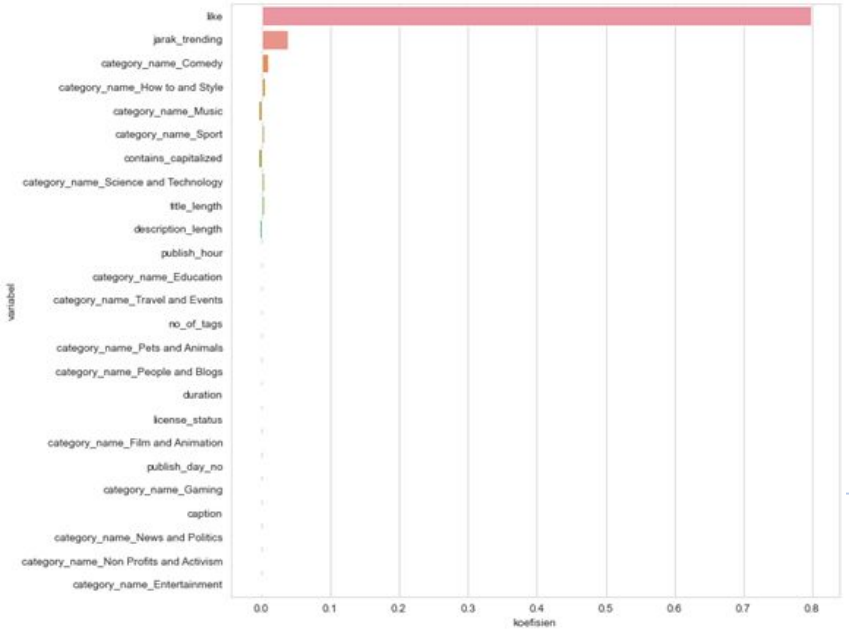
```
X_train  :  (42744, 25)
X_test   :  (10686, 25)
y_train  :  (42744,)
y_test   :  (10686,)
```

04

MODELLING

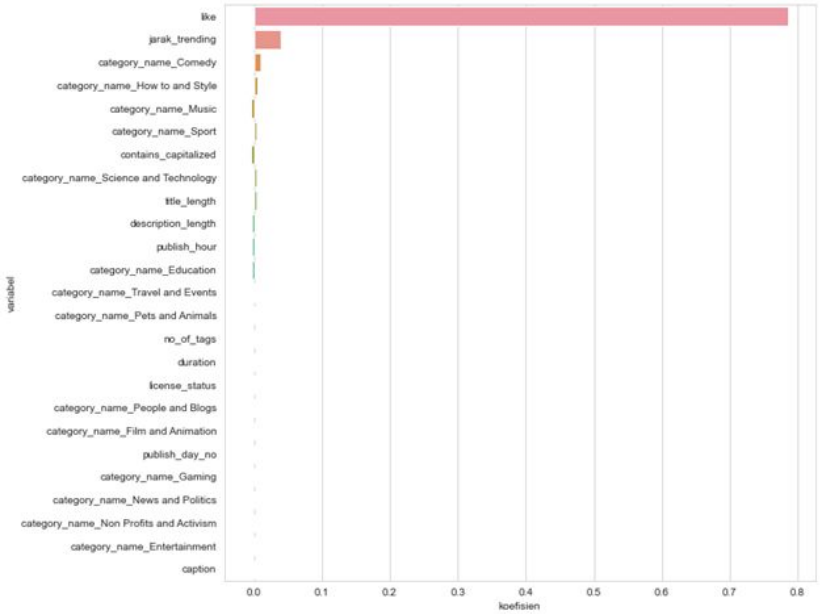
Linear Regression

	variabel	koefisien
0	duration	-0.000703
1	caption	-0.000198
2	license_status	-0.000633
3	like	0.798652
4	publish_day_no	-0.000429
5	publish_hour	-0.002024
6	jarak_trending	0.038245
7	description_length	-0.002225
8	title_length	0.003316
9	no_of_tags	0.001287
10	contains_capitalized	-0.003906
11	category_name_Comedy	0.009049
12	category_name_Education	-0.001962
13	category_name_Entertainment	-0.000091
14	category_name_Film and Animation	0.000509
15	category_name_Gaming	0.000215
16	category_name_How to and Style	0.005764
17	category_name_Music	-0.004528
18	category_name_News and Politics	0.000183
19	category_name_Non Profits and Activism	-0.000151
20	category_name_People and Blogs	-0.000721
21	category_name_Pets and Animals	0.001244
22	category_name_Science and Technology	0.003354
23	category_name_Sport	0.004113
24	category_name_Travel and Events	-0.001661



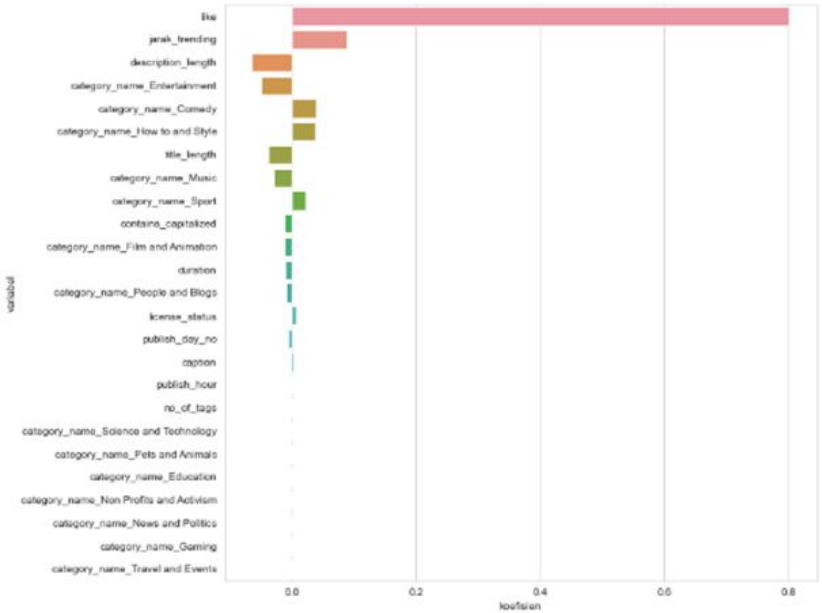
Ridge Regression

	variabel	koefisien
0	duration	-0.000970
1	caption	-0.000017
2	license_status	-0.000711
3	like	0.786304
4	publish_day_no	-0.000423
5	publish_hour	-0.001997
6	jarak_trending	0.039049
7	description_length	-0.002157
8	title_length	0.002910
9	no_of_tags	0.001199
10	contains_capitalized	-0.003857
11	category_name_Comedy	0.009131
12	category_name_Education	-0.001972
13	category_name_Entertainment	-0.000026
14	category_name_Film and Animation	0.000533
15	category_name_Gaming	0.000261
16	category_name_How to and Style	0.005839
17	category_name_Music	-0.004356
18	category_name_News and Politics	0.000196
19	category_name_Non Profits and Activism	-0.000076
20	category_name_People and Blogs	-0.000676
21	category_name_Pets and Animals	0.001233
22	category_name_Science and Technology	0.003316
23	category_name_Sport	0.004175
24	category_name_Travel and Events	-0.001668

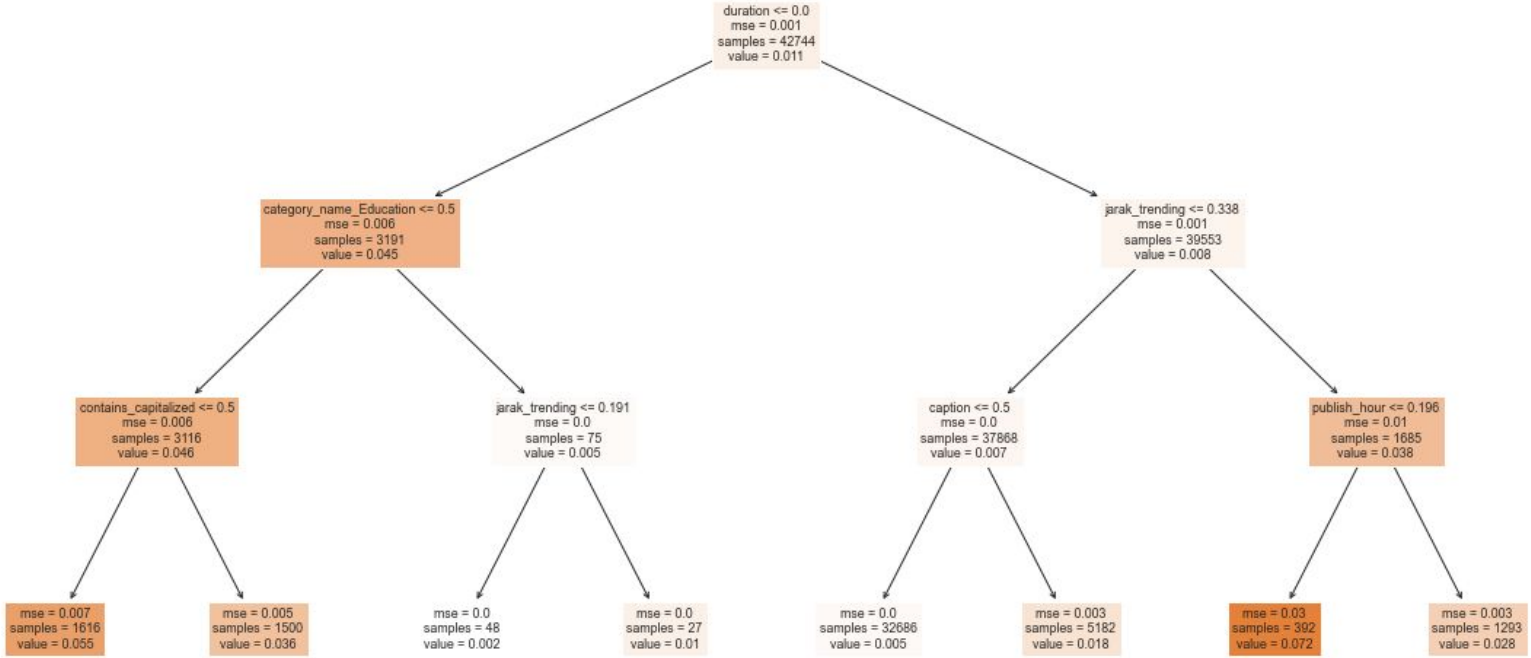


Support Vector Regression

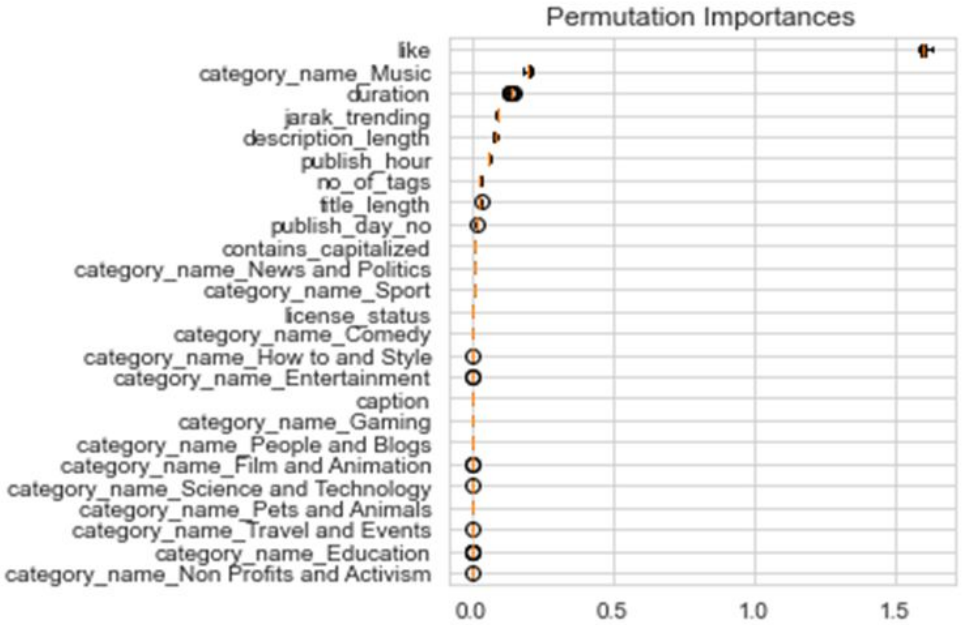
	variabel	koefisien
0	duration	-0.011382
1	caption	0.002502
2	license_status	0.007243
3	like	0.801257
4	publish_day_no	-0.005315
5	publish_hour	-0.002275
6	jarak_trending	0.088600
7	description_length	-0.065335
8	title_length	-0.037186
9	no_of_tags	-0.002234
10	contains_capitalized	-0.012490
11	category_name_Comedy	0.038756
12	category_name_Education	0.000000
13	category_name_Entertainment	-0.049053
14	category_name_Film and Animation	-0.011602
15	category_name_Gaming	0.000000
16	category_name_How to and Style	0.037392
17	category_name_Music	-0.028717
18	category_name_News and Politics	0.000000
19	category_name_Non Profits and Activism	0.000000
20	category_name_People and Blogs	-0.008443
21	category_name_Pets and Animals	0.000000
22	category_name_Science and Technology	0.000000
23	category_name_Sport	0.021668
24	category_name_Travel and Events	0.000000



Regression Tree



Regression Tree



05

EVALUASI

Perbandingan RMSE, R2, dan R2 Adj

	Linear Regression	Ridge Regression	SVR	Regression Tree
RMSE	0.0171	0.0170	0.34	0.0064
R2	77,9%	77,7%	12,2%	96,8%
R2 Adj	77,8%	77,7&	12,1%	96,8%

Regression tree adalah model terbaik karena menghasilkan RMSE paling kecil dan R2 Adjusted yang paling tinggi

06

KESIMPULAN DAN REKOMENDASI

Kesimpulan

1. Karakteristik dari data dapat dilihat pada bagian eksplorasi data
2. Variabel-variabel paling berpengaruh dilihat dengan melihat besar koefisien masing-masing dan perbandingan antar nilai koefisien
3. Model dipilih dalam tahap evaluasi dengan mempertimbangkan nilai RMSE, R2, dan Adjusted R2

Rekomendasi

1. Mempertimbangkan **jumlah *like*, kategori video, durasi, dan umur video** saat memasangkan iklan pada Youtube. Keempat variabel ini merupakan empat variabel yang paling berpengaruh terhadap jumlah *view* menurut model *regression tree* yang digunakan.

2. Berdasarkan eksplorasi yang dilakukan, sebaiknya prodi mengunggah video yang kontennya mengandung unsur **hiburan dan musik**. Selain itu, disarankan video diunggah di hari **Jum'at** pada rentang waktu **9.00 - 13.00 WIB**

3. Melakukan **pemodelan ulang** menggunakan lima variabel yang paling berpengaruh terhadap *view*. Hal ini dilakukan dengan tujuan untuk menghasilkan model yang lebih akurat dan presisi sehingga hasil model tidak lagi berpengaruh terhadap observasi yang digunakan sebagai *input*. Dengan kata lain, **model yang dihasilkan mampu menghasilkan *output* yang akurat dan presisi terlepas dari besar variansi *input* yang digunakan dalam model.**

TERIMA KASIH