Homework Solution - Regression

Team:

1. Avinska Valeria Gunawan
2. Dwirany Dian Audry Sinaga
3. Ichwan
4. Mega Dwi Astuti
5. Talenta Nanda

Submission:

1. Report: <https://docs.google.com/document/d/1ibu_Qxf3L7-9OVUCkqVuseKqSdDnsa4q/edit?usp=sharing&ouid=114571533115997718070&rtpof=true&sd=true>
2. Notebook: <https://colab.research.google.com/drive/1K06mRKAhJayU0WpuLT2OEIH8n_TaSwBm?usp=sharing>

# Pre-EDA Data Preprocessing

Saat mengerjakan tugas ini, dilakukan data preprocessing di awal sebelum melakukan EDA, agar pada saat EDA bisa melakukan eksplorasi ke hal-hal yang mendukung penyelesaian masalah. Adapun proses yang dilakukan adalah:

1. Load data ‘youtube\_statistics.xlsx” dan melihat df.info() untuk mengetahui jumlah baris dan juga jumlah kolom. Hasil: Jumlah baris 36791 dan jumlah kolom 18 kolom dengan tipe data: bool (3 kolom), datetime64[ns] (1 kolom), int64 (8 kolom) dan object(6 kolom)
2. Dikarenakan jumlah data/baris yang sangat banyak, hal pertama yang dilakukan adalah check duplikat agar proses selanjutnya lebih fokus pada data-data yang akan dipakai saja. Cek duplikat dilakukan 2 tahap:
   1. Melakukan check terhadap semua kolom dan diperoleh 4229 baris terduplikat dan semua didrop dari datase sehingga total baris saat ini 32562.
   2. Melakukan check terhadap kolom ***title, channel\_tittle,*** dan ***publish\_date*** karena jika video memiliki judul, channel dan publish date yang sama diasumsikan bahwa video tersebut adalah video yang sama. Maka dilakukan drop kembali dengan hasil akhir baris/data yang digunakan adalah 16493 baris.
3. Dilakukan pemisahan 18 kolom menjadi 2 jenis data yaitu numerical dan categorical untuk memudahkan melihat kolom-kolom tersebut.
   1. Numerical: ***category\_id, views, likes, dislikes, comment\_count, no\_tags, desc\_len, len\_title***
   2. Categorical: ***trending\_date, title, channel\_title, publish\_time, tags, comment\_disabled, rating\_disabled, video\_error\_or\_removed, description, publish\_date***
4. Dengan melihat dari df[cat].describe() dan df[nums].describe() ditemukan ada beberapa kolom yang dapat didelete karena memiliki jumlah kategori/variasi yang sangat banyak yaitu ***title, tags, description*** tetapi tidak dapat dilakukan pengolahan kembali seperti pengelompokan.
5. Dilakukan pengecekan missing value dimana tidak terdapat missing value dari hasil pengolahan yang dilakukan diawal.
6. Dilakukan pengecekan outliers dan ditemukan hampir seluruh data numerical memiliki outliers kecuali ***len\_title***. Akan tetapi, pada bagian ini tidak dilakukan handling outliers dikarenakan data outliers sangat banyak, tetapi masih bernilai valid sesuai dengan jenis datanya. Sebagai contoh: outliers pada data view masih valid karena jumlah penonton sebuah video masih memungkinkan untuk bernilai sangat tinggi

# Feature Engineering

Dilakukan feature engineering terlebih dahulu untuk memudahkan melihat EDA pada bagian feature/data kategorikal yang memiliki banyak variasi seperti:

1. Mengubah semua feature categorical yang masih memiliki beragam data type menjadi bertipe data ***category***.
2. Melakukan segmentasi pada data views dan membagi menjadi 3 bagian yaitu ***low, mid*** dan ***high*** berdasarkan percentile.
3. Melakukan segmentasi pada publish\_time dengan membagi menjadi 5 bagian yaitu ***morning, noon, afternoon, night, early morning***.

* Morning : 4-10 am
* Noon : 10 am - 2 pm
* Afternoon : 2 pm - 6 pm
* Night : 6 pm - 12 am
* Early morning : 12 - 4 am

1. Membagi trending dan publish date menjadi 2 kolom terpisah yang terdiri dari:
   * + **Bulan-Tahun**
     + **Day**

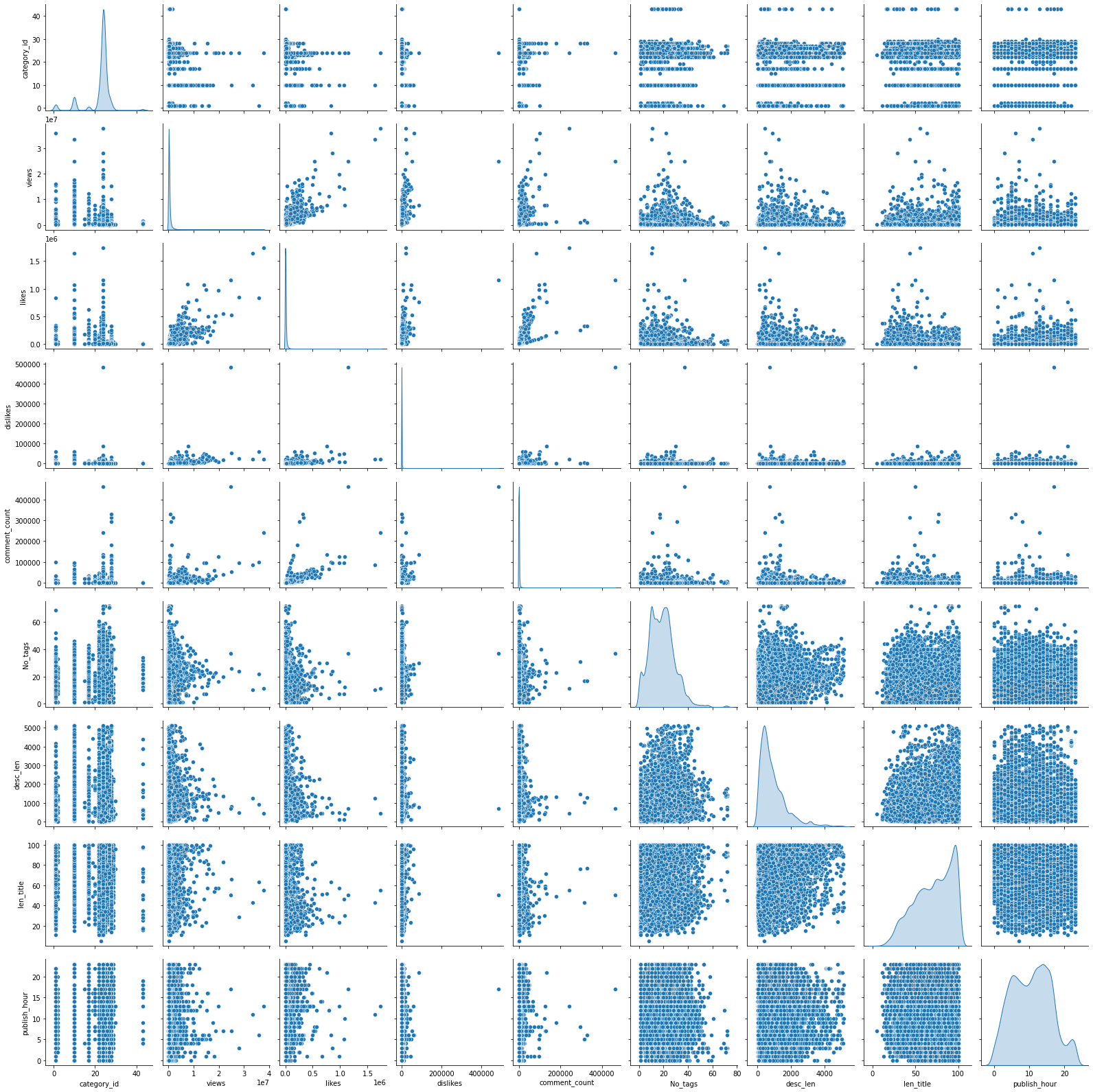
# 

# EDA

Dari sumber yang ada ditemukan bahwa standar yang membuat sebuah video trending di youtube adalah:

1. Durasi menonton
2. Rata-rata persentase views
3. Durasi video
4. Persentase penonton yang melihat video lalu keluar
5. Menonton ulang
6. Jumlah likes, dislikes
7. Jumlah viewers unik
8. Jumlah views per viewers unik
9. Darimana audience menonton (termasuk dari luar youtube)

Lalu dilakukan pair plot untuk melihat korelasi fitur numerical dan target(views)



Dari pairplot dan informasi yang diperoleh terkait hal-hal yang mempengaruhi views, maka dilakukan drop column pada column yang dianggap tidak berhubungan dengan views. Column yang didrop adalah ***no\_tags, ratings\_disabled, video\_error\_or\_removed, desc\_len,*** dan ***len\_title.***

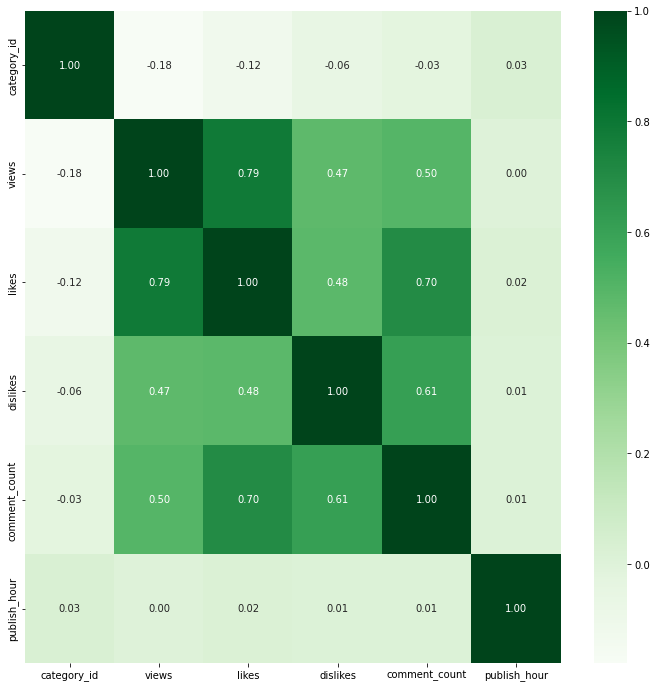
Yang diperoleh dari EDA adalah:

1. Dari data categorical dapat dilihat bahwa:

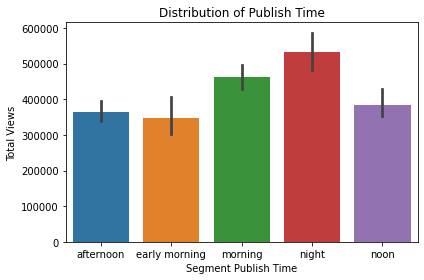
| **trending\_date** | **channel\_title** | **comments\_disabled** | **publish\_date** | **segment\_views** | **segment\_publish\_time** | **trending\_month** | **trending\_day** | **publish\_month** | **publish\_day** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 16493 | 16493 | 16493 | 16493 | 16493 | 16493 | 16493 | 16493 | 16493 |
| **unique** | 205 | 1390 | 2 | 219 | 3 | 5 | 8 | 31 | 9 |
| **top** | 2017-11-14 00:00:00 | Study IQ education | False | 2018-01-02 00:00:00 | mid | morning | December-2017 | 14 | December-2017 |
| **freq** | 192 | 172 | 15909 | 122 | 8246 | 5327 | 2851 | 774 | 2813 |

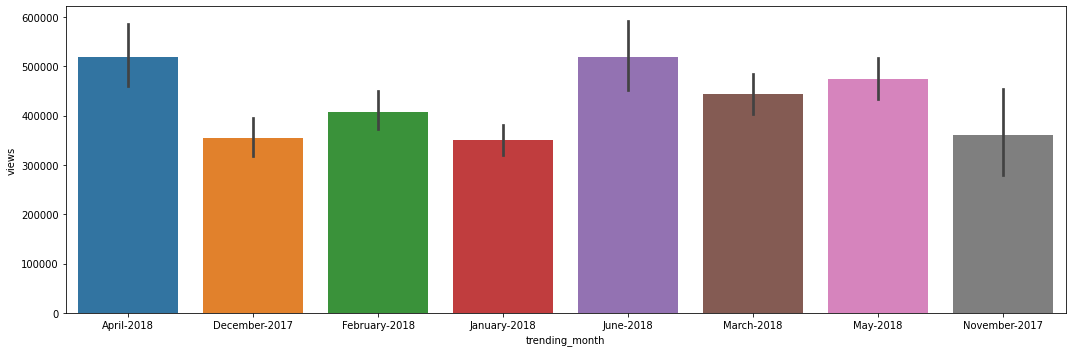
* 1. Video paling banyak di publish pada bulan Desember 2017 dan dari semua tahun paling banyak dipublish pada tanggal 16
  2. Video paling banyak trending pada bulan December 2017 dan dari semua tahun paling banyak trending pada tanggal 14
  3. Channel\_title paling tinggi frekuensinya adalah Study IQ Education dengan frekuensi 172
  4. Publish\_time paling tinggi yaitu pada pagi hari (morning) dengan frekuensi 5327
  5. Paling banyak video yang tidak melakukan disable comment dengan frekuensi 15909
  6. Video paling banyak berada pada kategori views “high” dengan frekuensi 6662

1. Dari heatmap terlihat bahwa views terlihat berkorelasi dengan likes dan comment\_count, dimana melalui heatmap terlihat views dan likes nilai korelasinya 0,79 serta views dan comment\_count nilai korelasinya 0,50.



1. Dari barplot dengan views, dapat dilihat views paling tinggi berada pada segment malam hari (night), trending month pada April-2018 dan June-2018. Sedangkan dari barplot trending day masih belum bisa diambil kesimpulan karena setiap bulan memiliki pola yang berbeda.





1. Dari kategori dapat dilihat bahwa viewer yang tinggi viewsnya adalah video dengan kategori 1, 10, 24 dan 17.

# Data Preprocessing

1. Sebelum melakukan modeling dilakukan preprocessing kembali untuk mengetahui publish time dan trending time termasuk pada jenis weekend atau tidak. Sehingga terdapat 4 kolom tambahan:

* ***publish\_day\_name***: berisi nama hari publish
* ***trending\_day\_name***: berisi nama hari trending
* ***is\_weekend***: bernilai 1 jika video dipublish pada weekend dan 0 jika dipublish pada weekdays
* ***is\_weekend2***: bernilai 1 jika video trending pada weekend dan 0 jika trending pada weekdays

1. Dilakukan feature encoding untuk beberapa kolom yaitu kolom ***publish\_day\_name, trending\_day\_name,*** dan ***segment\_publish\_time***
2. Dilakukan normalisasi untuk mendukung modeling pada semua fitur kecuali fitur target views.

# Modeling

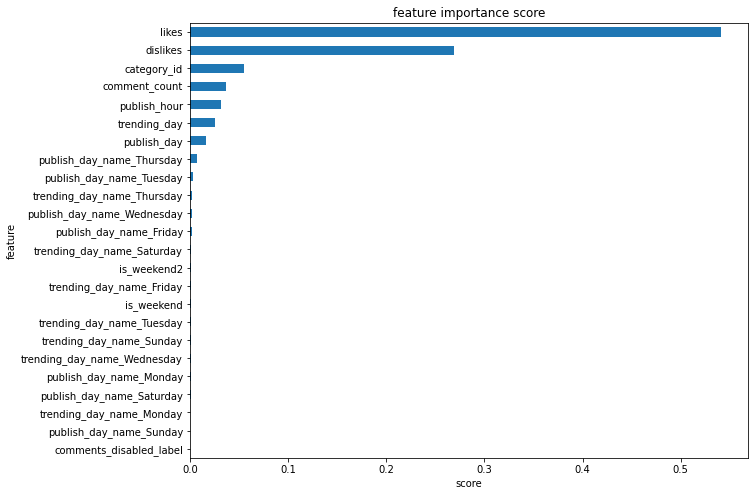
Tuliskan hasil dari modeling yang dilakukan. Jelaskan bagaimana performa RMSE dan R2 nya. Jika melakukan eksperimen lebih dari 1 algoritma atau parameter, deskripsikan bagaimana performa masing-masing, dan pilih salah satu model terbaik. Jelaskan mengapa menggunakan model tersebut.

Modeling dilakukan lebih dari 1 kali algoritma dengan hasil sebagai berikut:

| ***No*** | ***Algoritma*** | ***RMSE*** | ***R2*** |
| --- | --- | --- | --- |
| 1 | Linear Regression | 770235,08 | 0,58 |
| 2 | Ridge | 806435,40 | 0,54 |
| 3 | Ridge - Hyperparameter Tuning (Randomized Search) | 776337.24 | 0.57 |
| 4 | Ridge - Hyperparameter Tuning (Grid Search) | 776275.39 | 0.57 |
| 5 | Lasso | 770269.54 | 0.58 |
| 6 | Lasso - Hyperparameter Tuning (Randomized Search) | 770236.11 | 0.58 |
| 7 | ElasticNet | 1187071.26 | 0.00 |
| 8 | ElasticNet - Hyperparameter Tuning (Randomized Search) | 772601.44 | 0.58 |
| 9 | Decision Tree | 859279.88 | 0.48 |
| **10** | **Random Forest** | **687689.89** | **0.67** |
| 11 | Support Vector Regressor | 1215075.06 | -0.04 |
| 12 | Random Forest - Hyperparameter Tuning (Randomized Search) | 692592.28 | 0.66 |

Hyperparameter Tuning dengan Randomized Search dan Grid Search dinilai tidak jauh berbeda, dengan waktu proses yang lebih lama dalam penggunaan Grid Search, sehingga kedepannya kami menggunakan Randomized Search untuk menghemat waktu proses.

Berdasarkan hasil pemodelan dan evaluasi yang dilakukan terhadap nilai RMSE dan R2 score, didapatkan metode pemodelan yang paling baik menggunakan Random Forest dengan nilai RMSE 687689.89 dan nilai R2 score 0.67. Berdasarkan hasil hyperparameter tuning terhadap Random Forest, didapatkan 5 feature dengan importance score tertinggi untuk pemodelan ini secara berurutan yakni likes, dislikes, category\_id, comment\_count, dan publish\_hour.



# Summary

Dari modeling yang dilakukan dan feature importance yang diperoleh dapat diambil beberapa rekomendasi bisnis berdasarkan fitur-fitur penting yaitu:

* Dikarenakan likes dan comment berpengaruh ke views, maka direkomendasikan agar video yang diupload mengajak penonton untuk tidak lupa memberikan like dan comment.
* Berdasarkan hasil analisis EDA dengan menggunakan bar plot untuk melihat hubungan antara publish time dan viewers, direkomendasikan untuk mengupload video pada waktu malam dan pagi hari, karena waktu tersebut memiliki penonton lebih banyak dibandingkan waktu lainnya.
* Berdasarkan hasil analisis EDA, direkomendasikan untuk mengupload video dengan jenis video kategori (data diambil dari kaggle):
  + Kategori 1: Film & Animation
  + Kategori 10: Music
  + Kategori 24: Entertainment
  + Kategori 17: Sports

# Appendix

Kontribusi pengerjaan homework:

1. Data Pre-processing & EDA: Avinska Valeria Gunawan & Ichwan
2. Feature Engineering & Normalization: Mega Dwi Astuti
3. Modelling & Evaluation: Talenta Nanda & Dwirany Dian Audry Sinaga
4. Laporan: Semua