Homework Solution - Regression

I Gusti Ngurah Agung Hari Vijaya Kusuma Batch 57

August 3, 2025

Submission Links

- Repository: github.com/AgungHari

1 Pendahuluan

1.1 Latar Belakang

Youtube adalah platform berbagi video yang sangat populer di seluruh dunia. Dengan jutaan video yang diunggah setiap hari, penting bagi pembuat konten untuk memahami bagaimana cara meningkatkan visibilitas dan daya tarik video mereka. Salah satu cara untuk mencapai hal ini adalah dengan menganalisis metadata dari video yang telah diunggah sebelumnya.

Dalam tugas ini, kita akan menganalisis metadata dari video-video yang telah diunggah ke Youtube. Metadata ini mencakup berbagai informasi seperti judul, deskripsi, tag, dan jumlah penonton. Dengan menganalisis metadata ini, kita dapat mengidentifikasi pola-pola yang dapat membantu pembuat konten dalam meningkatkan performa video mereka.

1.2 Tujuan

Tujuan dari tugas ini adalah diantara lain untuk:

- Menganalisis metadata dari video Youtube untuk memahami faktor-faktor yang mempengaruhi jumlah penonton.
- Mengembangkan model prediksi yang dapat digunakan untuk memperkirakan jumlah penonton berdasarkan metadata video.
- Menerapkan regresi linier untuk memprediksi jumlah penonton video berdasarkan fitur-fitur yang tersedia dalam metadata.
- Membandingkan performa model regresi linier dengan model lain yang mungkin lebih kompleks.

1.3 Batasan Masalah atau Ruang Lingkup

Batasan masalah dalam tugas ini mencakup:

- Menggunakan dataset yang diberikan oleh rakamin.
- Menerapkan regresi linier sebagai metode utama untuk prediksi, meskipun model lain juga akan dieksplorasi.

1.4 Manfaat

Manfaat dari tugas ini adalah membuat model regresi linier yang dapat digunakan untuk memperkirakan jumlah penonton video Youtube berdasarkan metadata.

2 Tinjauan Pustaka

2.1 EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) adalah proses analisis data yang bertujuan untuk memahami struktur, pola, dan hubungan dalam dataset sebelum menerapkan model statistik atau machine learning. EDA melibatkan visualisasi data, statistik deskriptif, dan identifikasi anomali atau outlier. Proses ini penting untuk mendapatkan wawasan awal tentang data dan membantu dalam pengambilan keputusan selanjutnya.

2.2 Regresi Linier

Regresi linier adalah metode statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel independen (fitur) dengan variabel dependen (target). Model ini mengasumsikan bahwa hubungan antara variabel-variabel tersebut dapat direpresentasikan sebagai garis lurus. Regresi linier sering digunakan dalam analisis data untuk prediksi dan inferensi, serta merupakan dasar bagi banyak algoritma machine learning lainnya.

Agar dapat memahami bagaimana regresi linier bekerja, kita perlu memahami rumus dasar dari regresi linier. Regresi linier sederhana melibatkan satu variabel independen, sedangkan regresi linier berganda melibatkan beberapa variabel independen. Dalam tugas ini, kita akan fokus pada regresi linier berganda untuk memprediksi jumlah penonton video Youtube berdasarkan metadata yang tersedia.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{1}$$

Di mana:

- y adalah variabel dependen (target).
- β_0 adalah intercept (nilai awal ketika semua variabel independen bernilai nol).
- $\beta_1, \beta_2, ..., \beta_n$ adalah koefisien regresi yang menunjukkan pengaruh masing-masing variabel independen terhadap variabel dependen.
- $x_1, x_2, ..., x_n$ adalah variabel independen (fitur).
- ϵ adalah error term yang mencakup variasi yang tidak dijelaskan oleh model.

Dengan kemajuan teknologi dan ketersediaan pustaka machine learning yang kuat seperti scikit-learn, TensorFlow, dan PyTorch, penerapan regresi linier dalam model machine learning menjadi lebih mudah dan efisien. Pustaka-pustaka ini menyediakan fungsi-fungsi yang memungkinkan pengguna untuk dengan cepat membangun, melatih, dan mengevaluasi model regresi linier tanpa harus mengimplementasikan algoritma dari awal.

2.3 Metadata Youtube

Metadata Youtube mencakup berbagai informasi yang terkait dengan video, seperti judul, deskripsi, tag, kategori, dan statistik penonton. Metadata ini sangat penting karena membantu algoritma Youtube dalam merekomendasikan video kepada pengguna dan mempengaruhi visibilitas video di platform. Dengan menganalisis metadata, kita dapat mengidentifikasi faktor-faktor yang berkontribusi terhadap popularitas video dan mengembangkan strategi untuk meningkatkan performa konten.

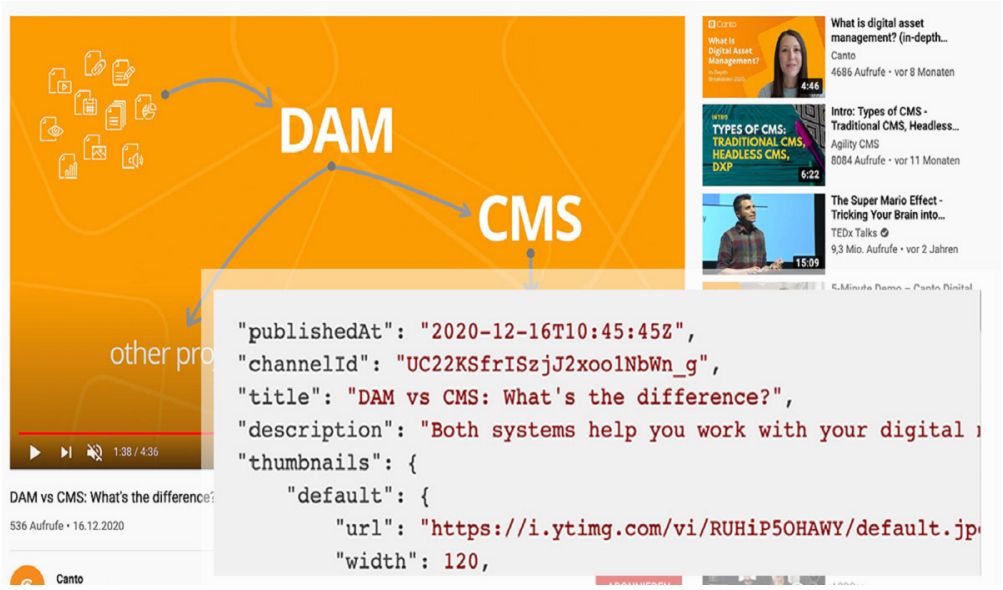

Figure 1: Contoh Metadata Youtube

2.4 RMSE (Root Mean Square Error)

Root Mean Square Error (RMSE) adalah metrik evaluasi yang digunakan untuk mengukur seberapa baik model prediksi dalam memprediksi nilai-nilai numerik. RMSE menghitung akar kuadrat dari rata-rata kuadrat selisih antara nilai yang diprediksi dan nilai aktual. Metrik ini memberikan gambaran tentang seberapa besar kesalahan prediksi model, dengan semakin kecil nilai RMSE menunjukkan performa model yang lebih baik.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

Di mana:

- n adalah jumlah data.
- y_i adalah nilai aktual.
- $\hat{y}_i$ adalah nilai yang diprediksi oleh model.

2.5 R^2 (Koefisien Determinasi)

Koefisien Determinasi (R^2) adalah metrik yang digunakan untuk mengukur seberapa baik model regresi menjelaskan variasi dalam data. Nilai R^2 berkisar antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan bahwa model mampu menjelaskan proporsi yang lebih besar dari variasi dalam data. Metrik ini sering digunakan untuk mengevaluasi performa model regresi dan membandingkan model yang berbeda.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

Di mana:

- n adalah jumlah data.
- y_i adalah nilai aktual.
- $\hat{y}_i$ adalah nilai yang diprediksi oleh model.
- $\bar{y}$ adalah rata-rata dari nilai aktual.

2.6 Modeling

Modeling dalam konteks machine learning adalah proses membangun model matematis atau statistik yang dapat digunakan untuk membuat prediksi atau mengambil keputusan berdasarkan data. Proses ini melibatkan pemilihan algoritma, pelatihan model dengan data, dan evaluasi performa model menggunakan metrik yang relevan. Dalam tugas ini, kita akan fokus pada penerapan regresi linier sebagai metode modeling utama untuk memprediksi jumlah penonton video Youtube berdasarkan metadata yang tersedia.

2.7 Scikit-learn

Scikit-learn adalah pustaka machine learning yang populer di Python yang menyediakan berbagai algoritma dan alat untuk analisis data dan modeling. Pustaka ini menawarkan antarmuka yang sederhana dan konsisten, sehingga memudahkan pengguna untuk menerapkan berbagai algoritma machine learning, termasuk regresi linier, klasifikasi, clustering, dan lain-lain. Scikit-learn juga menyediakan fungsi-fungsi untuk preprocessing data, evaluasi model, dan validasi silang, menjadikannya pilihan yang ideal untuk proyek-proyek machine learning.

2.8 Matplotlib dan Seaborn

Matplotlib dan Seaborn adalah pustaka visualisasi data yang populer di Python. Matplotlib menyediakan fungsi dasar untuk membuat berbagai jenis grafik, seperti garis, batang, dan sebar, sedangkan Seaborn adalah ekstensi dari Matplotlib yang menawarkan antarmuka yang lebih tinggi dan lebih mudah digunakan untuk visualisasi statistik. Keduanya sangat berguna dalam proses EDA (Exploratory Data Analysis) untuk memahami struktur dan pola dalam data melalui visualisasi yang informatif.

2.9 Pandas

Pandas adalah pustaka Python yang menyediakan struktur data dan fungsi untuk manipulasi dan analisis data. Pustaka ini menawarkan DataFrame, yang merupakan struktur data tabular yang memungkinkan pengguna untuk dengan mudah mengakses, memanipulasi, dan menganalisis data. Pandas sangat berguna dalam proses EDA (Exploratory Data Analysis) karena menyediakan berbagai fungsi untuk membersihkan, mengolah, dan menganalisis data secara efisien.

3 Desain dan Implementasi

Tugas ini dilakukan sesuai dengan desain sistem berikut beserta implementasinya. Desain sistem adalah konsep dari pembuatan dan perancangan infrastruktur dan kemudian diwujudkan dalam bentuk alur yang harus dikerjakan

3.1 Deskripsi sistem

Pada tugas ini kita akan melakukan analisis data dan membangun model regresi linier untuk memprediksi jumlah penonton video Youtube berdasarkan metadata yang tersedia. Proses ini melibatkan beberapa langkah, mulai dari eksplorasi data hingga evaluasi model. Desain sistem ini mencakup langkah-langkah yang akan dijabarkan dalam Gambar berikut.

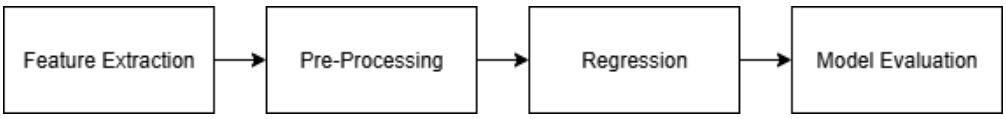

Figure 2: Blok Diagram Sistem

3.2 Feature Extraction

Feature extraction adalah proses penting dalam machine learning yang bertujuan untuk mengidentifikasi dan memilih fitur-fitur yang relevan dari data yang tersedia. Dalam konteks tugas ini, kita akan melakukan feature extraction pada metadata video Youtube untuk mendapatkan fitur-fitur yang akan digunakan dalam model regresi linier.

3.2.1 Deskripsi Dataset

Dataset yang digunakan dalam tugas ini adalah dataset dalam format .xslx yang berisi metadata dari video Youtube. Dataset ini mencakup berbagai informasi seperti judul, deskripsi, tag, kategori, dan statistik penonton. Setiap baris dalam dataset mewakili satu video, dan kolom-kolomnya berisi fitur-fitur yang relevan untuk analisis.

4 Summary dan Appendix

4.1 Summary

Simpulkan secara singkat hasil pengerjaan teman-teman. Jika ada rekomendasi bisnis yang diberikan akan lebih baik (nilai plus).

4.2 Appendix

Tuliskan kontribusi pengerjaan masing-masing anggota tim di bagian ini.

Tuliskan juga kesulitan-kesulitan dalam pengerjaan tugas ini di bagian ini (jika ada).