Homework Solution - Regression

Team: **Smile Project**

1. Piter Anjas (Ketua)
2. Ali Imron Nasrulloh
3. Davin Sabil Rizqullah
4. Fajar Akbardipura
5. M. Ilham Maulidi
6. Putu Nidia
7. Regita Afifah
8. Saia Mazaya Fatin

Submission:

1. Report: (https://docs.google.com/document/d/11cTqE0RgqNjcQsQMk2PR2qxi7PCzru1d7UYzuPtvxX0/edit?usp=sharing)
2. Notebook: (https://colab.research.google.com/drive/1drYV-iW_xGd76yF57SqQ43M4uSgTYY0i?usp=sharing)

Simple Exploratory Data Analysis

- Semua tipe data dari setiap kolom sudah sesuai, begitu juga dengan nama kolom dan isinya juga sudah sesuai. Terdapat 36791 baris data, dengan jumlah attribut 18.
- Description memiliki nilai kosong.
- Sebaran data feature likes, dislikes, comment_count, desc_len, dan views terlihat miring (selisih mean & median cukup besar)
- Pada grafik boxplot di atas bahwa fitur view, likes, dislikes, comment_count, o_tags, dan desc_len memiliki banyak outlier sehingga diperlukan transformasi logaritmik untuk fitur tersebut.
- Pada sebaran data fitur numerik terdapat beberapa fitur yang bersifat skewed negatif dan fitur len_title yang bersifat skewed positif.
- Terdapat 3 fitur yang memiliki korelasi positif kuat terhadap views, yaitu fitur likes, dislikes, dan comment_count

Data Preprocessing

- Menghapus data duplicate
- Menghapus data kosong
- Menghapus outlier

Feature Engineering

- Melakukan Log Transformation sebagai langkah untuk menghapus outlier dan untuk membuat distribusi data menjadi normal/hampir normal, hal ini dilakukan karena terdapat beberapa fitur yang memiliki sebaran data yang miring.
- Melakukan Normalization agar skala setiap fitur numerik mempunyai skala yang sama dan diharapkan dapat mempermudah proses pembelajaran data model machine learning yang akan dibuat.

Modeling

Hasil dari evaluasi modeling adalah sebagai berikut:

- Dari evaluasi **Model Linear Regression**, didapatkan bahwa nilai RMSE dan R^2 tidak mengalami overfitting karena hasil evaluasi antara nilai testing dan training tidak berbeda jauh, yaitu:

  RMSE (test): 0.012151278519364527
  RMSE (train): 0.012253153657086189
  R^2 (test): 0.779929481352955
  R^2 (train): 0.7622276257844693

- Begitu juga dengan hasil setelah dilakukan Hyperparameter Tuning, yang menunjukkan hasil nilai evaluasi nilai testing dan training tidak berbeda jauh. Performanya tidak bertambah karena model awal tidak overfitting.

  RMSE (test): 0.012151267295625778
  RMSE (train): 0.012253154383132344
  R^2 (test): 0.7799298878966554
  R^2 (train): 0.7622275976066257

- Model lain yang kami gunakan adalah model **Random Forest** dan **Model Lasso**:
  - Hasil dari Model Random Forest menunjukkan adanya kecenderungan overfitting, namun menurut kami masih dapat ditoleransi dalam batas kewajaran normal, dimana akurasi train memiliki nilai 99% dan akurasi test memiliki nilai 96,26%.
  - Hasil dari Model Lasso menunjukkan performa hasil yang jelek sehingga tidak cocok untuk digunakan. R^2 sebesar 0 berarti regresi Anda tidak lebih baik daripada mengambil nilai rata-rata, yaitu Anda tidak menggunakan informasi apa pun dari variabel lain. R^2 Negatif berarti kinerja Anda lebih buruk daripada nilai rata-rata.

Summary

- Dari hasil analisis dataset youtube_statistics.xlsx, bisa dilihat bahwa dataset perlu dianalisis dengan Exploratory Data Analysis. Dari hasil EDA, dilakukan data preprocessing diantaranya adalah menghapus data kosong, menghapus data duplikat, dan menghapus data outlier.
- Selain data preprocessing, dilakukan training model & prediksi data. Dari hasil training menggunakan Linear Regression Model, didapatkan kecenderungan overfitting yang sangat kecil, namun menurut kami masih dapat ditoleransi dalam batas kewajaran normal, dimana akurasi train memiliki nilai 99% dan akurasi test memiliki nilai 96,26%.
- Berdasarkan percobaan training model lain yang telah dilakukan, terdapat model yang baik untuk menentukan views video Youtube yaitu **Random Forest dengan RMSE 0.004, dan R^2 0.96 menjadikan Random Forest model terbaik.**

Appendix

Semua soal dikerjakan masing-masing terlebih dahulu kemudian dikerjakan dan didiskusikan bersama-sama.
Kesulitan yang dihadapi belum memahami materi secara keseluruhan dikarenakan waktu yang terbatas dan materi yang cukup padat, jadi agak kesulitan saat mengerjakan tugas.