

PROPOSAL PROYEK MACHINE LEARNING

Prediksi Nilai Rata-Rata Rating Buku



Mata Kuliah	Machine Learning (GRUP B)
Dosen Pengampu	<ul style="list-style-type: none">• Dr. Antonius Rachmat Chrismanto, S.Kom., M.Cs.• Gloria Virginia, S.Kom., MAI., Ph.D.• Dr. phil. Lucia Dwi Krisnawati, S.S.,M.A.• Dr., Ir. Sri Suwarno, M.Eng.
Kelompok	3
Anggota Kelompok	<ol style="list-style-type: none">1. 71231033 - Imel Rantetandung2. 71231052 - Gabriel Sachio Atmadjaja3. 71231058 - Michael Chandra Mahanaim
Deklarasi	Dengan ini kami menyatakan bahwa tugas ini merupakan hasil karya kelompok kami, tidak ada manipulasi data serta bukan merupakan plagiasi dari karya orang lain.

Topik / Judul

Topik / judul yang kami angkat berkaitan dengan dataset yang kami ambil adalah **“Prediksi Nilai Rata-Rata Rating Buku”** menggunakan algoritma machine learning berdasarkan fitur metadata buku.

Deskripsi

Rating buku merupakan indikator penting dalam menentukan kualitas dan popularitas sebuah buku di platform daring. Namun, rating tersebut seringkali baru tersedia setelah banyak pengguna memberikan ulasan. Oleh karena itu, diperlukan model machine learning yang dapat memprediksi nilai rata-rata rating buku berdasarkan karakteristik buku yang sudah diketahui sebelumnya seperti jumlah halaman, jumlah ulasan, jumlah pengguna yang memberikan rating, bahasa, dan penerbit.

Tujuan:

1. Membangun model machine learning yang mampu memprediksi nilai rata-rata rating buku.
2. Menganalisis hubungan antara fitur-fitur seperti jumlah halaman, jumlah rating, jumlah ulasan, dan penerbit terhadap rating buku.
3. Memberikan wawasan kepada penerbit atau penulis serta pembaca dalam menilai potensi kesuksesan buku sebelum dirilis.

Manfaat:

1. Membantu penerbit memperkirakan daya tarik buku baru.
2. Memberikan insight bagi penulis mengenai faktor-faktor yang mempengaruhi penilaian pembaca.
3. Dapat digunakan sebagai dasar dalam sistem rekomendasi buku

Dataset

Dataset yang kami gunakan diambil dari situs Kaggle <https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks> dengan judul “Goodreads-books”. Dataset ini berisi metadata ribuan buku yang diambil dari situs Goodreads. Dataset ini menggabungkan berbagai informasi seperti judul buku, nama penulis, jumlah halaman, jumlah rating, jumlah ulasan teks, tanggal publikasi, penerbit, serta bahasa buku.

Rencana / hasil EDA

Berikut ini rencana dan hasil setelah dilakukan EDA :

1. Statistik Deskriptif:
Menampilkan nilai rata-rata, median, standar deviasi dari kolom numerik (num_pages, ratings_count, text_reviews_count, average_rating).
2. Distribusi Rating:
Membuat histogram/boxplot untuk melihat sebaran nilai average_rating.
3. Korelasi antar variabel numerik:
Menggunakan *heatmap* untuk melihat hubungan antara jumlah halaman, jumlah rating, jumlah ulasan, dan rating rata-rata.

4. Analisis Penulis dan Penerbit:
Mengetahui siapa penulis/penerbit dengan rating rata-rata tertinggi.
5. Analisis Bahasa:
Mengetahui apakah bahasa tertentu cenderung memiliki rating lebih tinggi.
6. Outlier Detection:
Mendeteksi data ekstrim pada num_pages atau ratings_count.

Rencana Pre-Processing

Langkah-langkah yang akan dilakukan:

1. Handling Missing Values:
Mengecek dan menangani nilai kosong pada kolom publisher atau language_code.
2. Feature Encoding:
Mengubah data kategorikal (language_code, publisher, authors) menjadi bentuk numerik menggunakan Label Encoding atau One-Hot Encoding.
3. Feature Engineering:
 - a. Ekstraksi tahun dari kolom publication_date.
 - b. Membuat fitur baru seperti log_ratings_count dan log_reviews_count untuk mengurangi skewness
4. Normalisasi/Standarisasi:
 - a. Melakukan *scaling* pada fitur numerik seperti num_pages, ratings_count, dan text_reviews_count.
5. Split Data:
 - a. Memisahkan data menjadi data latih (80%) dan data uji (20%).

Rencana Metode ML

Model yang akan dibandingkan:

1. Linear Regression – sebagai model baseline.
2. Random Forest Regressor – untuk menangani non-linearitas antar fitur.
3. Gradient Boosting / XGBoost – untuk meningkatkan akurasi prediksi dengan ensemble method.

Tujuan dari penggunaan beberapa model adalah mencari model dengan performa terbaik untuk prediksi nilai rating.

Rencana Metode Evaluasi

Karena targetnya adalah nilai numerik (regresi), metrik evaluasi yang digunakan:

1. MAE (Mean Absolute Error): Mengukur rata-rata kesalahan absolut prediksi.
2. RMSE (Root Mean Squared Error): Mengukur kesalahan kuadrat rata-rata, sensitif terhadap outlier.
3. R^2 (Coefficient of Determination): Mengukur seberapa baik model menjelaskan variasi data target.

Model terbaik akan dipilih berdasarkan nilai RMSE paling kecil dan R^2 paling tinggi.

Kesimpulan

Proposal ini berfokus pada prediksi nilai rata-rata rating buku menggunakan data metadata buku. Dengan kombinasi EDA, preprocessing, dan model regresi (Linear Regression, Random Forest, dan XGBoost), diharapkan diperoleh model prediksi yang akurat serta wawasan mengenai faktor-faktor yang berpengaruh terhadap rating buku.