

# **PREDIKSI SKOR (RATING) ANIME BERDASARKAN *TYPE* DAN *GENRE***

*Disusun untuk memenuhi tugas pada mata kuliah:*

**Analitik Big Data (*Big Data Analytics*)**



*Ditulis oleh:*

**24917018-Eko Ikhwan Saputra**

**Program Studi Magister Informatika  
Fakultas Teknologi Industri  
Universitas Islam Indonesia  
2025**

# I

## PENDAHULUAN

### 1.1. Latar Belakang

Pasar anime telah berkembang sangat pesat, bahkan dalam satu tahun, dari \$26.000.000 pada tahun 2023 menjadi \$28.550.000 di tahun 2024 (The Business Research Company), pertumbuhan ini sebesar 9.8%. Tentunya perusahaan-perusahaan yang bergerak dalam industri anime melihat ini sebagai peluang besar, platform-platform *Over The Top* (OTT) adalah contohnya. Mereka menayangkan dan mendistribusikan anime untuk di tonton oleh pengguna. Sebagai perusahaan, mereka tentu ingin anime yang ditayangkan disukai oleh penggunanya, skor (rating) yang mencerminkan kualitas dan daya tarik dari anime merupakan indikator penting untuk menentukan keberhasilan suatu anime. Hal ini disebabkan karena skor adalah representasi seberapa besar pengguna (penonton) menikmati anime[1].

Skor anime sangat bervariasi, salah satu faktor yang menentukan adalah *genre*, biasanya setiap *genre* memiliki ciri khas yang berbeda dari *genre* lainnya[2]. Dalam dunia film, *genre* adalah salah satu atribut paling dasar untuk menentukan kesuksesan film[3], hal ini dapat juga diterapkan dalam anime. Selain *genre*, tipe dari penayangan anime juga menentukan skor, diantaranya adalah *Movie*, *TV Series* dan *Original Video Animation* (OVA). Setiap jenis penayangan memiliki karakteristik tersendiri, OVA biasanya memiliki skor lebih tinggi karena memiliki kualitas yang lebih baik dari *TV Series*[4] dan ditonton penggemar setia karena menceritakan cerita tambahan atau cerita alternatif[5]. Anime yang di distribusikan melalui *TV Series* biasanya bergantung pada popularitas manga dan hasil eksekusi[4], jika sesuai dengan ekspektasi penonton maka mendapatkan skor tinggi, dan juga sebaliknya. Sedangkan pada *movie* biasanya mendapat skor tinggi karena memiliki biaya dan waktu produksi yang jauh lebih tinggi[4], sehingga anime yang dihasilkan sangat memukau seperti *Your Name*.

Industri media dan hiburan sangat banyak menghasilkan data, baik dari riset, penjualan, basis pelanggan, file *log*, dll[6]. Termasuk data-data *genre* dan tipe dari anime yang tersedia di internet, salah satu platform yang menyediakan adalah *My Anime List*, situs ini adalah platform basis data anime paling populer di dunia. Teknologi *Big Data* bisa dipadukan dengan data-data tersebut untuk membantu pengambilan keputusan. Data

*genre* dan juga tipe dapat digunakan untuk membangun sistem yang bisa memprediksi skor dari anime. Tugas ini menggunakan 2 dataset

Tugas ini berfokus kepada pengembangan model prediksi skor anime berdasarkan *genre* dan tipe menggunakan teknologi *big data*, khususnya adalah *tools PySpark*. Model prediksi seperti ini dapat digunakan berbagai *enterprise* di industri anime, tetapi tugas ini memfokuskan penggunaan hasil prediksi pada platform OTT seperti *Crunchyroll*, *Netflix*, *Disney+*, *Amazon Prime*, dll. Prediksi skor anime dapat digunakan sebagai pertimbangan untuk menentukan apakah suatu anime mencapai target yang akan dicapai, hal ini berguna saat pengambilan keputusan terkait lisensi, promosi dan penayangan anime.

## **1.2. Pertanyaan**

Tugas ini bertujuan untuk dapat menjawab pertanyaan-pertanyaan berikut.

1. Apakah bisa membangun model prediksi skor anime berdasarkan variabel *genre* dan *type* menggunakan teknologi *PySpark*?
2. Seberapa akurat model prediksi skor anime yang dikembangkan dengan dataset *My Anime List*.
3. Bagaimana hasil prediksi skor anime dapat membantu platform *Over The Top* (OTT) seperti *Crunchyroll* dan *Netflix* untuk pengambilan keputusan?

## **1.3. Objektif (Tujuan)**

Untuk mendukung pertanyaan yang dipaparkan sebelumnya, berikut adalah tujuan dari pengerjaan tugas berikut ini.

1. Mengembangkan model prediksi skor anime menggunakan variabel *genre* dan *type* dengan mengutamakan penggunaan teknologi *PySpark*.
2. Mengevaluasi performa model prediksi menggunakan metrik *Root Mean Squared Error* (RMSE).
3. Memberikan rekomendasi yang dapat digunakan oleh platform *Over The Top* (OTT) berdasarkan hasil prediksi skor.

4. Mengilustrasikan atau memaparkan bagaimana memanfaatkan teknologi *PySpark* untuk membantu pengambilan keputusan strategis pada industri media dan hiburan, khususnya anime.

#### 1.4. Ruang Lingkup dan Batasan

##### 1. Ruang Lingkup

- Penelitian ini menggunakan 2 dataset kaggle yang di *crawling* dari situs *My Anime List*. Fokus pada variabel *genre* dan *type* sebagai **X** dan variabel *score* sebagai **y**.
- Model prediksi yang dirancang diharapkan dapat membantu platform *Over The Top* (OTT) dalam menentukan apakah sebuah anime dapat mencapai target atau layak dijadikan prioritas lisensi, promosi ataupun penayangan.
- Teknologi *PySpark* digunakan sebagai *tools* utama, tetapi tidak semua tahap pengerjaan menggunakan *PySpark*.

##### 2. Batasan pengerjaan tugas ini adalah.

- Tidak seluruh tahapan pengerjaan menggunakan *PySpark*.
- Tugas ini hanya menggunakan variabel *genre*, *type* dan *score*, tidak mempertimbangkan variabel lain pada dataset.
- Model yang dikembangkan hanya menggunakan algoritma regresi sederhana tanpa eksplorasi berbagai algoritma lain yang mungkin lebih kompleks.
- Model diuji dalam lingkungan teknis dan belum diterapkan dalam kasus nyata di industri.

#### 1.5. Dataset

Tugas ini menggunakan 2 dataset yang didapat dari *Kaggle*, dimana kedua dataset ini mengumpulkan dari dari situs *My Anime List*. Berikut adalah dataset yang digunakan.

##### 1. *Top 10,000 Anime (Popularity Index of 2024) – Dataset 1*

- Ada 16739 anime yang terdata pada dataset.
- Memiliki 12 kolom (*uid*, *title*, *synopsis*, *genre*, *aired*, *episodes*, *members*, *popularity*, *ranked*, *score*, *img\_url*, dan *link*).

- Ada beberapa kolom yang memiliki *missing value* yaitu *synopsis* (975), *episodes* (706), *ranked* (3212), *score* (579), dan *img\_url* (180).
- *Missing value* pada kolom *score* yang akan menjadi target (variabel y) akan ditangani pada bab selanjutnya.
- *Genre* paling banyak atau paling sering diproduksi adalah *Music*.

## 2. *Anime Dataset – Dataset 2*

- Ada 10000 anime yang terdata.
- Memiliki 6 kolom (*Name*, *Rating*, *Ranked*, *Popularity*, *Members* dan *Type*).
- Seluruh kolom pada dataset ini tidak memiliki *missing value*.
- *Type* anime paling banyak adalah *TV*.

## II TAHAPAN PEKERJAAN

### 2.1. Pengumpulan Data

*Dataset* diperoleh dari situs *Kaggle*, tugas ini menggunakan 2 *dataset* seperti yang telah dijabarkan pada bab sebelumnya. Kedua *dataset* ini digabungkan menjadi satu *dataset* saja. Sebelum disatukan, kolom *title* pada *dataset 1* dan kolom *Name* pada *dataset 2* diubah dengan nama *anime\_title*, selain itu kedua kolom ini dilakukan operasi *lower case*, hal ini dilakukan untuk menyamakan format sehingga kedua *dataset* dapat disatukan berdasarkan judul anime nya. Penghapusan data duplika dilakukan setelah penggabungan *dataset*, kemudian beberapa kolom dihapus yaitu *popularity*, *Members*, dan *Ranked*. Nilai kosong pada kolom *score* akan digantikan oleh nilai *Rating*, hal ini dilakukan karena nilai *Rating* merujuk kepada hal yang sama pada *score*, yaitu angka penilaian penonton terhadap suatu anime dalam rentan 1 – 10. Kolom *score* juga kemudian dipastikan berformat *float*. Tahap-tahap ini belum memanfaatkan *PySpark*, tetapi *pandas*.

### 2.2. Pre-processing

Setelah data digabungkan, data perlu dipersiapkan sebelum proses membuat model prediksi. Karena data sudah melewati beberapa proses pembersihan pada tahap sebelumnya, pada tahap ini hanya memecah kolom *genre*, format atau bentuk nilai dari kolom ini merupakan *list* dari *genre-genre* di setiap anime, sehingga harus dipisahkan menjadi baris baru. Tahap ini dilakukan dengan menggunakan *PySpark*. Teknik *One-Hot-Encoding* diterapkan pada kolom *genre* dan *type*, hal ini berguna untuk memetakan atau mengkonversi setiap *genre* dan *type* kedalam bentuk numerik sehingga dapat dipelajari oleh algoritma.

### 2.3. Modelling

Pada tahap ini adalah membuat *pipeline modelling* untuk setiap algoritma, dimana algoritma yang digunakan pada tugas ini adalah *Linear Regression*, *Random Forest Regressor* dan *Gradient-Boosted Trees Regressor*. Sebelum model dilatih *dataset* dibagi menjadi *training set* dan *test set* dengan perbandingan 70:30, dimana kolom *genre* dan *type* adalah X sedangkan kolom *score* adalah y. Setelah itu maka setiap

algoritma dilatih sehingga menghasilkan sebuah model. Model yang telah dilatih ini dievaluasi untuk melihat seberapa baik performanya dalam memprediksi skor anime. Metrik yang digunakan untuk mengevaluasi model adalah *Root Mean Squared Error* (RMSE). Model *Linear Regression* dilatih dengan parameter *default*, model *Random Forest Regressor* dibangun dengan parameter *numTrees* = 300, dan model *Gradient Boosted Trees Regressor* dibangun dengan parameter *maxIter*=100. Seluruh tahapan *modelling* ini menggunakan *tools PySpark*.

#### **2.4. Deployment Sederhana**

*Deployment* sederhana ini dibuat hanya untuk melihat contoh pengaplikasian prediksi skor anime, sehingga dapat dibayangkan penggunaannya. Tahap ini tidak menggunakan *tools* yang rumit atau *deploy* dalam bentuk web, hanya user interaktif sederhana saja. Dengan ini pengguna dapat memasukkan *input* dan melihat *output* dari prediksi. *Tools* yang digunakan adalah *widgets* dan *display*.

### III HASIL DAN PEMBAHASAN

#### 3.1. Hasil

Total anime yang ada pada kedua *dataset* adalah 26739 ( $16739 + 10000$ ), setelah *dataset* dijadikan satu berdasarkan judul anime, yaitu dengan kolom *anime\_title*, total anime menjadi 8568, lalu dilakukan operasi penghapusan duplikasi data, sehingga total anime setelah data duplikat dihapus yaitu 7353. Sekarang *dataset* baru kita namakan *final\_df*, beberapa kolom pada *final\_df* memiliki *missing value*, yaitu *synopsis*(215), *episodes*(235), *ranked*(257), *score* (255), dan *img\_url*(17). Karena tugas ini hanya berfokus pada variabel *genre* dan *Type* sebagai X serta *score* sebagai y, maka *missing value* yang ditangani hanyalah kolom *score*. Nilai-nilai kosong pada kolom *score* akan digantikan dengan nilai dari kolo *Rating*, kolom ini memiliki *value* yang sama yaitu penilaian penonton terhadap suatu anime dalam rentang 1 – 10.

Setelah itu dilakukan *explode* pada kolom *genre*, sehingga setiap *genre* pada setiap anime terpisah kedalam baris baru, karena kolom *genre* berisi *list* dari *genre* setiap anime. Setelah itu dilakukan pemisahan *training data* dan *test data*, *training data* digunakan untuk melatih model atau untuk model belajar, sedangkan *test data* digunakan untuk mengevaluasi performa dari setiap model yang telah dibangun. Data dibagi 70:30, 70% *data training* dan 30% *data testing*. Setelah data dibagi maka tahap *modelling* dilakukan, hasil dari *modelling* dapat dilihat pada tabel 1.

Tabel 1. Performa model

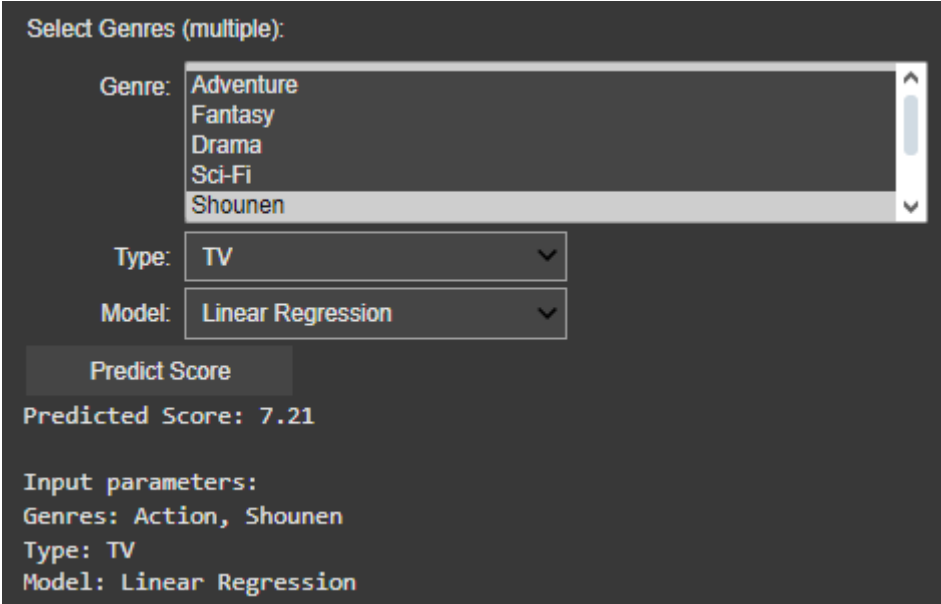
Algoritma (Model)	RMSE
Linear Regression	0.6180
Random Forest Regressor	0.6226
Gradient-Boosted Trees Regressor	0.6184

Tidak ada perbedaan nilai RMSE yang jauh diantara ketiga model, model *Linear Regression* mendapatkan performa terbaik dengan nilai RMSE terkecil, yaitu 0.6180, *Gradient-Booster Trees Regressor* tidak jauh dengan nilai RMSE 0.6184, sedangkan *Random Forest Regressor* dengan nilai RMSE 0.6226. Nilai RMSE di angka tersebut



tentu masih belum cukup baik untuk memprediksi skor anime, karena masih cukup tinggi, semakin kecil nilai RMSE maka semakin baik performa model tersebut.

Model-model yang telah dilatih sudah dapat digunakan untuk memprediksi skor anime, dalam tugas ini saya mencoba membuat *deployment* sederhana untuk melihat gambaran ketika model ini dikembangkan lebih lanjut dan digunakan. Pada *deployment* sederhana ini pengguna dapat memilih kombinasi beberapa *genre* anime dan juga tipe distribusi atau penayangan anime nya, lalu pengguna juga dapat memilih model atau algoritma mana yang ingin digunakan, setelah itu pengguna bisa mendapatkan hasil prediksi skor anime nya. Hasil dari *deployment* sederhana menggunakan *widgets* dan *display* dapat dilihat pada gambar 1.



Select Genres (multiple):

Genre: Adventure  
Fantasy  
Drama  
Sci-Fi  
Shounen

Type: TV

Model: Linear Regression

Predict Score

Predicted Score: 7.21

Input parameters:  
Genres: Action, Shounen  
Type: TV  
Model: Linear Regression

Gambar 1. Hasil *deployment*

Gambar 1 menampilkan contoh penggunaan prediksi skor anime, anime dengan *genre* Action dan Shounen ditayangkan pada penayangan TV diprediksi mendapatkan skor 7.21 oleh model *Linear Regression*.

### 3.2. Pembahasan

Melalui tahapan *Data Science Cycle* atau *Artificial Intelligence Lifecycle* dan sebagainya model prediksi anime berdasarkan *genre* dan *type* dengan *dataset* dari platform *My Anime List* dapat dibangun menggunakan teknologi *PySpark*. Pada tugas ini *PySpark* digunakan hampir pada keseluruhan proses pengerjaan.

Performa model yang dikembangkan masih belum maksimal, hal ini bisa terjadi karena tidak adanya *hyperparameter tuning*, kurangnya variabel dan algoritma *machine learning* tradisional. Algoritma yang lebih kompleks mungkin akan menghasilkan model dengan performa yang lebih baik lagi.

Prediksi skor anime dapat digunakan oleh platform *Over The Top* (OTT) dalam menambah wawasan untuk menilai sebuah anime memenuhi kualitas atau target yang ingin dicapai. Sebagai contoh, anime dengan prediksi skor diatas rata-rata dapat di prioritaskan untuk dipromosikan, atau didapatkan lisensi dan penayangannya. Tentunya prediksi skor anime bukan menjadi satu-satunya tolak ukur, karena platform OTT memiliki banyak data variabel-variabel lain, tetapi setidaknya prediksi skor anime dapat menambah wawasan atau sudut pandang dalam pengambilan keputusan ini.

## IV PENUTUP

### 4.1. Kesimpulan

Objektif pada tugas ini telah dapat dicapai dengan selesainya model prediksi anime. Ada beberapa hal yang dapat disimpulkan, yaitu sebagai berikut.

1. Teknologi *PySpark* dapat digunakan untuk menangani data berskala besar mulai dari proses *pre-processing* hingga membangun model prediksi.
2. Model regresi yang dikembangkan mampu untuk memprediksi skor anime walaupun hasil evaluasi nilai *Root Mean Squared Error* masih belum cukup baik.
3. Variabel *genre* dan *Type* dapat digunakan untuk memprediksi skor anime.
4. Hasil prediksi model dapat digunakan platform *Over The Top* (OTT) untuk menambah sudut pandang atau wawasan dalam mengambil keputusan terkait lisensi, promosi dan penayangan anime.
5. Penggabungan dua dataset sangat berguna, dalam kasus ini nilai pada kolom *score* yang kosong dapat diganti dengan nilai dari kolom *Rating*.

### 4.2. Saran

Saran untuk tugas selanjutnya adalah pengembangan dari model ini, yaitu sebagai berikut.

1. Menambah variabel lain untuk menentukan skor anime.
2. Menggunakan algoritma yang lebih kompleks.
3. Menambahkan data dari sumber selain *My Anime List*.
4. Membuat *dashboard visualisasi* yang relevan untuk membantu menambah sudut pandang lain selain prediksi skor anime sebelum pengambilan keputusan.

## REFERENSI

- [1] A. N. Varma dan K. Petluri, "Movie Recommender System using critic consensus," dalam *2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*, Mumbai, India: IEEE, Des 2021, hlm. 1–4. doi: 10.1109/ICAC353642.2021.9697196.
- [2] N. W. Ummah, D. Puspitasari, dan M. Aryadillah, "Struktur Formula Anime Genre Isekai Fantasi," *EAR*, vol. 2, no. 1, hlm. 32–43, Mar 2024, doi: 10.22146/ear.12216.
- [3] S. Sahu, R. Kumar, H. V. Long, dan P. M. Shafi, "Early-production stage prediction of movies success using K-fold hybrid deep ensemble learning model," *Multimed Tools Appl*, vol. 82, no. 3, hlm. 4031–4061, Jan 2023, doi: 10.1007/s11042-022-13448-0.
- [4] N. C. Khabib, "PERKEMBANGAN DAN PENGARUH ANIME DI INDONESIA, SERTA PENGUNAAN ANIME SEBAGI MEDIA PEMBELAJARAN," Diponegoro University, Semarang, 2022.
- [5] B. H. Ini, "Apa Itu OVA dalam Anime? Ini Penjelasannya yang Perlu Dipahami Otaku," kumparan. Diakses: 14 Januari 2025. [Daring]. Tersedia pada: <https://kumparan.com/berita-hari-ini/apa-itu-ova-dalam-anime-ini-penjasannya-yang-perlu-dipahami-otaku-1ygUsHF4f11>
- [6] H. Lippell, "Big Data in the Media and Entertainment Sectors," dalam *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*, J. M. Cavanillas, E. Curry, dan W. Wahlster, Ed., Cham: Springer International Publishing, 2016, hlm. 245–259. doi: 10.1007/978-3-319-21569-3\_14.

## **LAMPIRAN**

### **Dataset**

Anime Dataset : [Link Dataset 1](#)

Top 10,000 Anime (Popularity Index of 2024) : [Link Dataset 2](#)

### **GitHub (Kode Python)**

[Link Repository GitHub \(Kode\)](#)