

PREDIKSI PELUANG KESUKSESAN FILM DALAM PRA PRODUKSI MENGUNAKAN ALGORITMA DECISION TREE

Dhani Ariatmanto, Muhammad Ilham Arief

Program Studi Magister Teknik Informatika S2, Fakultas Ilmu Komputer
Universitas AMIKOM Yogyakarta, Jalan Padjajaran, Ring Road utara, Yogyakarta, Indonesia
dhaniari@amikom.ac.id, ilhamarieff@students.amikom.ac.id

ABSTRAK

Film merupakan salah satu hiburan yang populer didunia. Tingkat kesuksesan film tergantung dari jumlah penonton. Namun, banyaknya jumlah penonton berkaitan tidak hanya dari pemeran utama tapi juga plot cerita dan genre dari film tersebut. Pra produksi merupakan tahapan dari proses pembuatan Film. Dalam proses pra produksi terdapat ide dan konsep untuk pembuatan naskah. Penelitian ini memprediksi kuseksesan produksi film dari pengklasifikasian berdasar bahasa, negara, title years, imdb_score, movie_title, content_rating, director_name, budget, gross, genre, actor_name. Dataset yang digunakan bersifat public dari IMDB dan metode yang digunakan yaitu decision tree (DT). Tahapan dimulai dari pengumpulan data, pre-processing, klasifikasi dan terakhir pengujian model. Dari hasil eksperimen didapatkan tingkat akurasi yang lebih baik dari penelitian sebelumnya dengan akurasi sebesar 68%.

Kata kunci: *Film, Prediction, Decision Tree, Random Forest, Success*

1. PENDAHULUAN

Industri Film saat ini dan kedepannya akan terus berkembang. Memutuskan untuk berinvestasi diindustri film sangatlah beresiko. Suksesnya film bergantung jumlah penonton yang hadir dan ketertarikan akan aktor dan alur cerita. Beberapa film dianggap sukses berdasarkan apakah sudah masuk dalam box office global atau dilihat dari hasil rating IMDb, dimana setiap tahun ratusan film dibuat dan didistribusikan [1]. Ada film bagus dan ada juga film yang kurang menarik perhatian penonton. Oleh karena itu, sangatlah penting bagaimana rumah produksi film bisa menentukan kualitas mereka sebelum benar-benar film tersebut ditayangkan atau bagaimana rumah produksi menentukan genre film yang memuaskan untuk ditonton [2]. Sehingga menghindari kerugian financial yang besar.

Situs website IMDb merupakan wadah informasi populer terkait dengan ulasan dan reaksi penonton terhadap film yang ditonton. Karena keunggulannya, situs website IMDb adalah rumah bagi banyak data terkait film serta ulasan pengguna. Peringkat yang diberikan oleh IMDb diakui dengan baik oleh masyarakat umum, mencerminkan persetujuan penonton akan kualitas suatu film [3]. Data dari IMDb dapat digunakan sebagai dataset untuk melakukan klasifikasi film berdasarkan atribut-atributnya dan sebagai cara untuk dapat memprediksi kesuksesan sebuah karya rumah produksi film. Atribut-atribut yang terdapat dalam film dapat mencakup: genre film, aktor film, bahasa, warna, durasi film, negara, dan lain-lain yang dijadikan tolak ukur sutradara dan produser dalam memproduksi sebuah film. Dengan besarnya cakupan data yang tersebar di internet, dan berkembangnya kemampuan computer salah satunya machine learning, dapat dijadikan model algoritma dalam memprediksi tingkat kesuksesan produksi film.

Adanya model machine learning prediksi ini dapat membantu investor dan pemangku kepentingan dalam proses pengambilan keputusan.

Metode Decision Tree (DT), Rule based (RB), dan Random Forest (RF) adalah tiga metode yang dapat digunakan untuk mengkategorikan data [4]. Distribusi variabel prediktor bukanlah batasan pada salah satu dari tiga pendekatan nonparametrik. Metodologi pembelajaran mesin induktif yang efektif adalah metode Decision Tree terdiri dari diagram alur akar-simpul-cabang-daun yang secara efektif menyimpan data dengan menguji kualitas kumpulan data setiap simpul pohon. Cabang menunjukkan berbagai hasil yang mengarah ke simpul berikut, dan daun (atau simpul terminal) mewakili kelas ketika proses ini selesai. Pendekatan Decision Tree memiliki berbagai manfaat dibandingkan teknik klasifikasi yang lain seperti kemungkinan maksimum dan pengklasifikasi fungsi diskriminan linier [5].

Tujuan dari penelitian ini yaitu prediksi peluang kesuksesan film dalam pra produksi sebelum film itu di buat dilihat dari beberapa aspek dan menjadi bahan pertimbangan rumah produksi dalam pembuatan film di masa depan, dengan implementasi Algoritma Decision Tree sebagai alat ukur apakah tingkat akurasi dari penelitian ini sudah baik. Adapun dataset diambil dari Kaggle.com yaitu Dataset IMDb yang nantinya diolah untuk mengetahui prediksi peluang kesuksesan, Adapun Langkah-langkah dalam memproses dataset tersebut yaitu pertama yaitu dilakukan pengumpulan data, preprocessing data, klasifikasi, pengujian model. Hasil perbandingan algoritma Decision Tree dan Metode machine learning yang sudah digunakan yang nantinya menjadi tolak ukur peneliti dalam melihat tingkat akurasi.

2. TINJAUAN PUSTAKA

Adapun beberapa penelitian yang sudah membahas terkait dengan prediksi kesuksesan film. Dalam [2], memaparkan skema penelitian mengenai prediksi kesuksesan film menggunakan algoritma machine learning dengan metode-metode algoritma yaitu *random forest*, *G boost*, *adaboost classsifier*, SVM dan KNN. Penelitian mereka menganalisa keluaran dari setiap metode-metode algoritma kemudian dilakukan perbandingan. Penelitian mereka menggunakan dataset dari IMDb yang terdapat di situs website keagel sejumlah 5043 Film dengan 28 Atribut. Performa dari hasil eksperimen didapatkan tingkat akurasi yang berbeda-beda dimana random forest = 61%, G boost = 56 %, adaboost classsifier= 41 %, SVM= 45 % dan KNN= 44%. Namun, tingkat akurasi dari hasil prediksi mereka masih rendah dan masih dapat ditingkatkan.

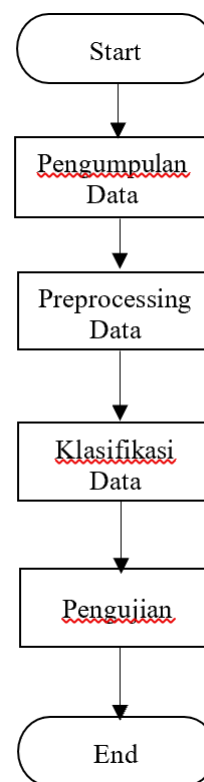
Dalam [6], penulis mengembangkan algoritma machine learning untuk meprediksi kesuksesan film dengan penerapan beberapa model antara lain: support vector machine(SVM), k-Nearest Neighbor (K-NN) dan Gaussian Naïve Bayes(GNB) untuk memprediksi kesuksesan film. Dataset yang digunakan didalam penelitian ini sejumlah 5000 film dengan pemilihan 11 atribut. Dari hasil eksperimen didapatkan tingkat akurasi yaitu SVM= 80.4%, K-NN= 81%, dan GNB= 85.4%. pada penelitian ini disimpulkan bahwa metode GNB merupakan algoritma machine learning yang bekerja lebih baik. Namun dalam penelitian ini tidak melibatkan penerapan metode algoritma decision tree.

Dalam [7], peneliti mempresentasikan algoritma ANN, Naïve bayes dan SVM untuk memprediksi kesuksesan film. Peneliti menggunakan dataset dari IMDb dimana yang digunakan dalam pengujian sebanyak 1500 data film. Dari hasil eksperimen didapatkan tingkat akurasi dari masing-masing metode yaitu ANN = 86%, Naïve bayes = 65%, dan SVM = 63%. Didapatkan hasil kesimpulan bahwa metode algoritma ANN menghasilkan tingkat akurasi yang paling tinggi dari metode Naïve bayes dan SVM.

Dalam penelitian ini penggunaan algoritma decision tree akan diterapkan dengan mengambil dataset dari IMDb. Dataset yang digunakan berjumlah 5043 data film. Dalam ujicoba eksperimen akan menggunakan 28 atribut. Hasil dari eksperimen akan di bandingkan dengan penelitian dari Rijul Dhir dan Anand Raj.

3. METODOLOGI

Tahapan dari metodologi penelitian yang akan dilakukan diantaranya: pengumpulan dataset dari IMDb, dilanjutkan dengan dilakukannya preprocessing data, pengklasifikasian data, dan untuk tahapan terakhir pengujian model, dimana nilai akurasi didapatkan. Diagram proses dari rancangan penelitian prediksi kesuksesan film digambarkan pada Gambar 1.



Gambar 1. Metodologi penelitian

3.1. Pengumpulan Dataset

Untuk dapat prediksi skor pada website IMDb secara efektif, kami akan mencoba mengidentifikasi atribut-atribut setiap dataset yang mempengaruhi. Dalam penelitian ini. 5043 Dataset Film IMDb diambil dari situs Kaggle.com. Ini memiliki 28 variabel untuk 5042 film dan 4906 poster film yang menjangkau 100 tahun dan 66 negara. Ada ribuan pemain dan aktris dan 2399 nama sutradara yang berbeda [2].

Dalam dataset yang digunakan dalam penelitian ini merupakan dataset publik yang pernah dilakukan penelitian sebelumnya oleh Rijul Dhir dan Anand Raj berjumlah 5043 film dan dengan 28 atribut dari <https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset> [2]. Detail dataset penelitian dapat dilihat pada Tabel 1 dan Tabel 2.

Tabel 1. Dataset Penelitian

Jenis Data	Jumlah Data
Film	5043
Atribut	28

Tabel 2. Jenis Data Atribut

No	Atribut
1	Color
2	Director_name
3	num_critc_for_reviews
4	duration
5	director_facebook_likes
6	actor_3_facebook_likes

No	Atribut
7	actor2_name
8	actor_1_facebook_likes
9	gross
10	genres
11	actor_1_name
12	movie_title
13	num_voted_users
14	cast_total_facebook_likes
15	actor_3_name
16	facenumber_in_poster
17	plot_keywords
18	movie_imdb_link
19	num_user_for_reviews
20	language
21	country
22	content_rating
23	budget
24	title_year
25	actor_2_facebook_likes
26	imdb_score
27	aspect_ratio
28	movie_facebook_likes

3.2. Pre-Processing Data

Pada tahap ini dilakukan pengolahan data yaitu preparation data dimana data diperbaiki dengan cara menghapus data yang tidak digunakan dan data yang bersifat null, juga menghapus 2 atribut yaitu color dan movie imdb link, dan menghapus data yang duplikat sehingga data yang sudah di preparation menjadi 4695 film dan 26 atribut, dapat dilihat pada Tabel 3.

Tabel 3. Data setelah Pre-processing

Jenis Data	Jumlah Data
Film	4695
Atribut	26

3.3. Klasifikasi Data

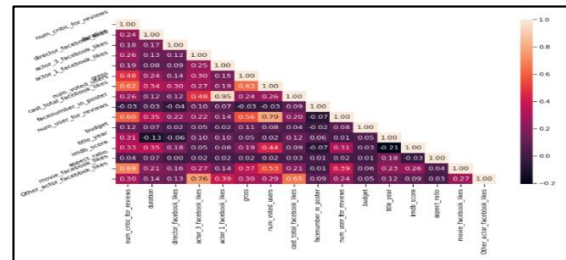
Dalam pengklasifikasian data ini dilakukan perbandingan data setiap atribut dimana data tertinggi di dalam kelasnya yang diambil sebagai bahan prediksi kesuksesan film adapun atribut yang pilih dalam penelitian yaitu antara lain Bahasa, Negara, Title Years, imdb_score, movie_title, content_rating, director_name, budget, Gross, Genre, actor_1_name yang hasilnya dapat dilihat pada Tabel 4.

Tabel 3. Klasifikasi Data

Nilai	Atribut	Data Atribut Tertinggi
4500	Bahasa	bahasa inggris
3568	Negara	USA
250	Title Years	tahun 2000an
12000	imdb_score	6.9
1.2	movie_title	The host
24000	content_rating	R (Restricted)
9.3	director_name	Frank darabont
1.3	budget	554.793.859
24000	Gross	7 imdb score
209	Genre	Drama
5.9	actor_1_name	Michael jai white

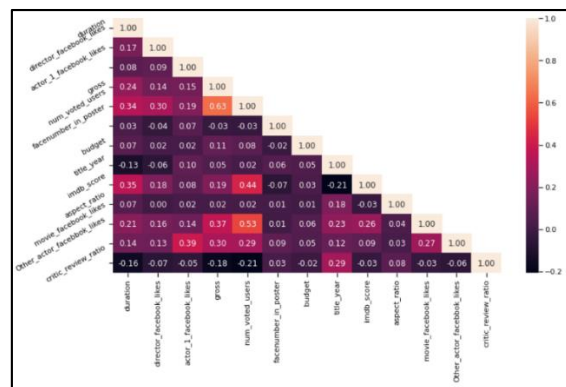
3.4. Pengujian

Dalam pengujian data pada penelitian ini, data yang berkorelasi seperti pada Gambar 2 akan di perbaiki.



Gambar 2. Data sebelum perbaikan korelasi

Melihat bahwa `cast_total_facebook_likes` dan `actor_1_facebook_like` sangat berkorelasi satu sama lain. Baik `aktor2` dan `aktor3` juga entah bagaimana berkorelasi dengan total. Jadi kami ingin diubah menjadi dua variabel: `actor_1_facebook_likes` dan `other_actors_facebook_likes`. Ada korelasi tinggi di antara `num_voted_users`, `num_user_for_reviews`, dan `num_critic_for_reviews`. Peneliti mempertahankan `num_voted_users` dan rasio `num_user_for_reviews` dan `num_critic_for_reviews`. Data sudah perbaikan korelasi dapat dilihat pada Gambar 3.



Gambar 3. Data setelah perbaikan

Sekarang kita dapat melihat tidak ada atribut yang tidak banyak berkorelasi satu sama lain. Decision Tree merupakan Teknik yang paling terkenal untuk mewakili pengklasifikasi untuk klasifikasi data atau biasa disebut pengklasifikasi pohon keputusan. pohon keputusan dari data yang dapat diakses, seperti pembelajaran mesin, pengenalan pola, dan statistik, telah dipelajari oleh beberapa akademisi dari berbagai profesi dan latar belakang. Penggunaan pengklasifikasi pohon keputusan telah disarankan dalam berbagai domain, termasuk analisis penyakit medis, kategorisasi teks, klasifikasi ponsel cerdas pengguna, gambar, dan banyak lagi. Pohon keputusan melakukan pendekatan dengan sangat rinci mengklasifikasi data [8].

4. HASIL DAN PEMBAHASAN

Dalam proses prediksi kesuksesan film berdasarkan dataset IMDb ini dengan menggunakan data awal sebesar 5043 film dan 28 atribut setelah dilakukan preprocessing data yaitu menghapus data yang memiliki nilai Null, dan menghapus atribut yang tidak masuk dalam pengklasifikasian yaitu color, dan movie imdb link, dan menghapus data yang duplikat dalam dataset sehingga didapatkan 4737 film dan 26 atribut.

Setelah itu dilakukan pengelompokkan data pada atribut yang memiliki nilai yang tinggi dalam kelasnya kemudian nilai tertinggi ini menjadi acuan dalam aspek pembuatan film yang sukses, untuk pengujian akurasi dilakukan perbandingan sebesar 20: 80 sehingga Evaluasi model ini akan menghasilkan accuracy, recall, precision, dan F1-Score sebesar 68%, 0.69, 0.57, dan 0.62 untuk nilai confusion matrixnya dapat dilihat pada Gambar 4 dan akurasi dilihat pada Tabel 4.

0	7	8	11	0
1	15	128	107	0
2	15	110	479	13
3	0	1	17	28
	0	1	2	3

Gambar 4. Confusion Matrix

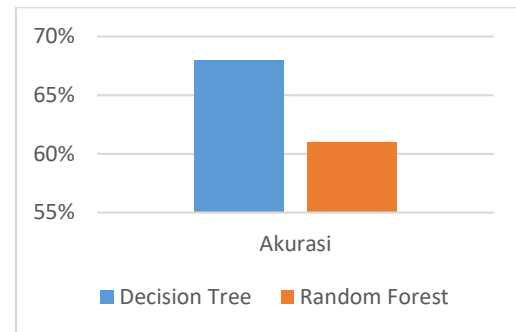
Tabel 4. Hasil

Decision Tree	Hasil
Akurasi	0.68
precision	0.69
recall	0.57
f1-score	0.62

Ketika membandingkan dengan penelitian sebelumnya yang dilakukan oleh [2] Rijul Dhir dan Anand Raj dengan dataset yang sama dengan Algoritma *Random Forest* sebagai algoritma yang nilai akurasi sebesar 61% maka pada penelitian ini mengalami peningkatan sebesar 7% dapat dilihat pada Tabel 5.

Tabel 5. Perbandingan hasil akurasi penelitian sebelumnya

Algoritma	Akurasi
Random forest	61%
Decision Tree	68%



Gambar 5. Perbandingan akurasi dengan penelitian sebelumnya

gambar 5. menggambarkan grafik perbandingan akurasi antara metode *Decision Tree* (DT) dan *Random Forest* (RF). Dimana metode DT dari eksperimen yang telah dilakukan mampu meningkatkan tingkat akurasi sebesar 7% dibandingkan dengan metode RF.

5. KESIMPULAN DAN SARAN

Berdasarkan dari penelitian yang dilakukan terkait prediksi kesuksesan film dengan penerapan algoritma *Decision Tree*. Dapat ditarik kesimpulan bahwa metode algoritma DT mengalami peningkatan akurasi sebesar 7 % dari penelitian sebelumnya dengan akurasi DT sebesar 68% dan RF sebesar 61%. Kedepan akan dilakukan penelitian dengan metode yang berbeda untuk meningkatkan tingkat akurasi menjadi lebih baik.

DAFTAR PUSTAKA

- [1] N. Quader, M. O. Gani, and Di. Chaki, "Performance evaluation of seven machine learning classification techniques for movie box office success prediction," in *3rd International Conference on Electrical Information and Communication Technology, EICT 2017*, 2018, vol. 2018-Janua, pp. 1–6, doi: 10.1109/EICT.2017.8275242.
- [2] R. Dhir and A. Raj, "Movie Success Prediction using Machine Learning Algorithms and their Comparison," in *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications*, 2018, pp. 385–390, doi: 10.1109/ICSCCC.2018.8703320.
- [3] J. Chen and M. Cui, "Analysing Gender Bias in IMDB Films Based on Social Networks," in *IOP Conference Series: Materials Science and Engineering*, Apr. 2020, vol. 806, no. 1, doi: 10.1088/1757-899X/806/1/012022.
- [4] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
- [5] T. M. Berhane *et al.*, "Decision-tree, rule-based, and random forest classification of high-

- resolution multispectral imagery for wetland mapping and inventory,” *Remote Sens.*, vol. 10, no. 4, Apr. 2018, doi: 10.3390/rs10040580.
- [6] Ankit, M. Lakshmi, K. A. Shastry, A. Sandilya, and R. Shekhar, “A comparative analysis of Machine Learning approaches for Movie Success Prediction,” in *Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud, ISMAC 2020*, Oct. 2020, pp. 684–689, doi: 10.1109/I-SMAC49090.2020.9243589.
- [7] R. A. Masrury, M. A. A. Saputra, A. Alamsyah, and M. A. S. Primantari, “A comparative study of Hollywood movie successfulness prediction model,” 2019, doi: 10.1109/ICoICT.2019.8835385.
- [8] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.