

Klasifikasi Kesuksesan Film Berdasarkan Karakteristik Produksi Menggunakan Random Forest

Garin Dani Saputra^{1*}, Elsa Oktawiyani², Aditama Aditama Nurani³, Alif Akbar Andriyansyah⁴, Sultan Syah Afriyadi⁵

¹Teknik Informatika/Ilmu Komputer

Universitas Duta Bangsa Surakarta

^{1*}230103061@mhs.udb.ac.id

²Teknik Informatika/Ilmu Komputer

Universitas Duta Bangsa Surakarta

²230103058@mhs.udb.ac.id

³Teknik Informatika/Ilmu Komputer

Universitas Duta Bangsa Surakarta

³230103049@mhs.udb.ac.id

⁴Teknik Informatika/Ilmu Komputer

Universitas Duta Bangsa Surakarta

⁴230103048@mhs.udb.ac.id

⁵Teknik Informatika/Ilmu Komputer

Universitas Duta Bangsa Surakarta

⁵230103078@mhs.udb.ac.id

Abstrak— Perkembangan industri perfilman global yang pesat memicu timbulnya risiko finansial yang tinggi bagi rumah produksi dan investor akibat ketidakpastian pasar. Kemampuan memproyeksikan potensi kesuksesan film sejak tahap pra-produksi menjadi instrumen manajemen risiko yang sangat krusial. Penelitian ini bertujuan membangun model klasifikasi kesuksesan film menggunakan algoritma *Random Forest Classifier* berdasarkan karakteristik produksinya. Siklus metodologi yang diadopsi adalah *Cross-Industry Standard Process for Data Mining* (CRISP-DM) yang mencakup enam tahapan terstruktur. Dataset yang digunakan bersumber dari *TMDB 5000 Movies*. Label target biner kesuksesan film didefinisikan berdasarkan pencapaian skor evaluasi rating akhir lebih dari tujuh koma nol. Model dilatih menggunakan konfigurasi parameter *n_estimators* sama dengan seratus dan *random_state* sama dengan empat puluh dua dengan melibatkan fitur utama berupa *budget*, *revenue*, *popularity*, *runtime*, genre populer, dan keterlibatan aktor kelas atas. Berdasarkan hasil pengujian, model *Random Forest* berhasil mencatatkan performa akurasi global yang tinggi sebesar 83,8%. Hasil analisis *feature importance* menunjukkan bahwa atribut durasi film (*runtime*) dan tingkat popularitas di masyarakat merupakan dua parameter paling dominan dalam memengaruhi keputusan percabangan pohon keputusan. Model terbaik diserialisasi ke dalam format .pkl dan diimplementasikan ke dalam *backend* aplikasi web berbasis *framework* Flask sebagai sistem pendukung keputusan praktis secara *real-time*.

Kata kunci— Data Mining, CRISP-DM, Klasifikasi, Kesuksesan Film, Random Forest, Flask, TMDB.

Abstract— The rapid growth of the global film industry has led to high financial risks for production houses and investors due to market uncertainty. The ability to project a movie's potential success during the pre-production phase serves as a crucial risk management tool. This study aims to develop a movie success classification model using the Random Forest Classifier algorithm based on production characteristics. The research methodology adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, encompassing six structured stages. The dataset is sourced from the TMDB 5000 Movies. The binary target label for movie success is defined based on achieving a final evaluation rating score of more than seven point zero. The model is trained using parameters *n_estimators* equals one hundred and *random_state* equals forty-two, utilizing key features such as budget, revenue, popularity, runtime, popular genres, and top-tier actors involvement. Experimental results demonstrate that the Random Forest model achieves a high global accuracy of 83.8%. Feature importance analysis indicates that movie duration (runtime) and public popularity are the two most dominant parameters influencing the decision tree branches. The best-performing model is serialized into a .pkl format and integrated into a web-based application backend using the Flask framework to provide a real-time, practical decision support system.

Keywords— Data Mining, CRISP-DM, Classification, Movie Success, Random Forest, Flask, TMDB.

I. PENDAHULUAN

Industri perfilman global terus mengalami pertumbuhan produksi yang signifikan, dengan jumlah film yang diproduksi pada tahun 2024 tercatat melampaui level sebelum pandemi [1]. Setiap film yang diproduksi melibatkan investasi biaya yang besar, mulai dari tahap pra-produksi, produksi, hingga distribusi dan promosi. Namun, besarnya biaya produksi tidak selalu berbanding

lurus dengan tingkat penerimaan penonton terhadap film tersebut [2]. Akibatnya, banyak film dengan budget tinggi justru mendapatkan penilaian rendah, sementara beberapa film dengan budget terbatas justru berhasil meraih apresiasi tinggi dari masyarakat.

Kondisi ini mendorong para analis, studio, dan investor untuk mengembangkan model yang mampu memprediksi potensi kesuksesan film

secara objektif sebelum film dirilis. Kemampuan untuk memproyeksikan penerimaan penonton sejak tahap pra-produksi menjadi instrumen manajemen risiko yang sangat krusial dalam industri kreatif. Kesuksesan film dalam penelitian ini didefinisikan berdasarkan rating penonton, yaitu film dikategorikan sukses apabila memperoleh rata-rata penilaian (vote average) minimal 7,0.

Penelitian sebelumnya telah menunjukkan efektivitas machine learning dalam memprediksi kesuksesan film. Aritmanto & Arief (2023) menerapkan algoritma Decision Tree untuk memprediksi peluang kesuksesan film pada tahap pra-produksi dan memperoleh akurasi sebesar 68% [3]. Pada konteks perfilman Indonesia, Athaya et al. (2024) menerapkan Jaringan Syaraf Tiruan untuk memprediksi kesuksesan film berdasarkan jumlah tiket terjual [4], sementara Sianipar et al. (2025) menunjukkan pentingnya analisis genre terhadap popularitas film [5].

Berbeda dengan penelitian sebelumnya yang sering kali menggunakan fitur pasca-rilis (seperti revenue atau jumlah vote) yang menyebabkan data leakage, penelitian ini secara ketat membatasi variabel input hanya pada karakteristik pra-produksi yang tersedia sebelum film diproduksi, seperti budget, durasi film (runtime), genre, dan keterlibatan aktor utama. Penggunaan algoritma Random Forest dipilih karena kemampuannya dalam menangani hubungan non-linear antar variabel serta ketahanannya terhadap overfitting dibandingkan pohon keputusan tunggal [6]. Selain itu, model ini diharapkan memberikan hasil yang lebih andal bagi rumah produksi dalam perencanaan strategis [7].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk: (1) membangun model klasifikasi Random Forest untuk memprediksi kesuksesan film berdasarkan karakteristik pra-produksi; (2) mengevaluasi performa model secara komprehensif menggunakan metrik accuracy, precision, recall, dan F1-score untuk menangani ketidakseimbangan data; serta (3) mengidentifikasi karakteristik produksi yang paling berpengaruh terhadap kesuksesan film melalui analisis feature importance.

II. METODOLOGI PENELITIAN

Penelitian ini mengadopsi siklus standar industri CRISP-DM yang terdiri dari 6 tahapan terstruktur untuk mengolah data dari bentuk mentah hingga diimplementasikan menjadi sebuah sistem pendukung keputusan [8].

A. Business Understanding

Tujuan utama dari tahap ini adalah menerjemahkan kebutuhan industri perfilman menjadi tujuan teknis *data mining*. Masalah utama yang dihadapi adalah tingginya risiko finansial dalam produksi film akibat minimnya alat ukur objektif di fase pra-produksi. Target teknis dari penelitian ini adalah membangun model klasifikasi biner yang mampu menentukan keputusan proyeksi apakah suatu proyek film akan masuk kategori sukses atau gagal berdasarkan karakteristik nilai produksi yang direncanakan.

B. Data Understanding

Dataset utama yang melandasi eksperimen ini bersumber dari data publik Kaggle (*TMDB 5000 Movies*). Karakteristik data dipahami dengan mengidentifikasi parameter-parameter kunci pembentuk performa film. Atribut intrinsik yang dianalisis meliputi judul film (title), biaya produksi (budget), total pendapatan grosir (revenue), tingkat popularitas di masyarakat (popularity), durasi penayangan (runtime), rating rata-rata (vote_average), jumlah penilai (vote_count), serta kategori genre utama (main_genre).

C. Data Preparation

Tahap penyiapan data merupakan tahapan krusial untuk menjamin kualitas data komputasi sebelum diumpankan ke dalam model [9]. Skenario penyiapan data meliputi tiga langkah utama:

1. Data Cleaning

Mengeliminasi entri data duplikat, menangani nilai kosong (*missing value*) pada atribut runtime, serta membuang seluruh baris film yang mencatatkan angka 0 pada kolom finansial budget maupun revenue karena dianggap tidak valid untuk analisis profitabilitas komersial.

2. Feature Engineering & Labeling

Mengekstrak array kompleks JSON pada data mentah menjadi satu variabel teks tunggal berupa

genre utama (*main_genre*) [10]. Selanjutnya, dilakukan pembentukan kolom target biner kesuksesan berdasarkan perolehan skor evaluasi akhir rating film ($\geq 7,0$).

3. Data Splitting

Dataset bersih hasil olahan akhir kemudian dibagi menggunakan metode pengacakan otomatis (*random split*) dengan proporsi 80% dialokasikan sebagai data latih (*training data*) untuk pembentukan cabang pohon dan 20% disisihkan sebagai data uji (*testing data*) untuk validasi objektif model [7].

D. Modeling

Pada tahap pemodelan, algoritma *Random Forest Classifier* diimplementasikan dengan menetapkan parameter *n_estimators* = 100 (menggabungkan 100 arsitektur *decision tree* independen) serta kunci acak *random_state* = 42. Model melatih data dengan membaca fitur masukan utama yang tersedia pada dataset bersih, yaitu: *budget*, *popularity*, *runtime*, *vote_average*, *vote_count*, *release_year*, dan *main_genre* yang telah ditransformasikan ke bentuk representasi numerik [11].

E. Evaluation

Skenario evaluasi diukur berdasarkan tingkat akurasi global (*Accuracy*), kemampuan tebakan presisi (*Precision*), sensitivitas temuan kembali (*Recall*), serta nilai rata-rata harmonis (*F1-Score*) dari hasil prediksi terhadap data uji [12]. Seluruh metrik tersebut dipetakan dan dihitung secara matematis menggunakan representasi tabel *Confusion Matrix* serta analisis kontribusi *Feature Importance* [13].

F. Deployment

Model *Random Forest* terbaik disimpan ke dalam format berkas biner .pkl menggunakan pustaka *pickle*. Model diintegrasikan ke dalam arsitektur aplikasi berbasis web menggunakan *framework* Flask. Tahap ini dirancang untuk menghasilkan antarmuka grafis formulir input parameter film komersial bagi pengguna akhir (*end-user*).

III. HASIL DAN PEMBAHASAN

A. Hasil Data Preparation

Rangkaian pembersihan data pada tahap *Data Preparation* berhasil menyeleksi sampel

data yang valid dari dataset mentah *TMDB 5000 Movies*. Proses eliminasi baris yang memiliki nilai kosong (*missing value*) atau tidak logis (seperti nilai anggaran atau pendapatan bernilai nol) menghasilkan sekumpulan data bersih yang siap digunakan untuk proses pemodelan. Pratinjau dari lima sampel teratas dataset hasil pemrosesan (*tmdb_5000_movies_CLEAN.csv*) disajikan pada Tabel 1.

Tabel 1. Pratinjau Dataset Hasil Pemrosesan

No	Judul	Budget (\$M)	Pop	Dur (Mnt)	Kelas (Y)
1	Avata	237.0	150.4	162	1
2	Pirate	300.0	139.0	169	0
3	Spectr	245.0	107.3	148	0

Keterangan: Nilai finansial dikonversi ke dalam jutaan USD (M). Kolom kelas Y merepresentasikan target biner (1 untuk rating $\geq 7,0$ dan 0 untuk rating $< 7,0$).

Tabel 1 menunjukkan bahwa parameter produksi telah ditransformasikan ke dalam struktur numerik yang ringkas. Konversi ini krusial agar komputasi algoritma berjalan efisien dalam batas ruang memori. Setelah memastikan seluruh baris data bersih, dataset kemudian dibagi menggunakan teknik *random split* berasio 80:20 untuk memisahkan data pelatihan (*training*) dan data pengujian (*testing*).

B. Hasil Pemodelan dan Evaluasi

Setelah model dilatih menggunakan data latih, pengujian performa dilakukan menggunakan sisa sampel data uji. Hasil pengujian riil komputer menunjukkan bahwa model *Random Forest* mencatatkan performa yang sangat tinggi dengan Model Accuracy mencapai 83,8%.

Guna melihat ketepatan klasifikasi model secara mendalam berdasarkan rancangan evaluasi CRISP-DM, representasi penyebaran keputusan dipetakan secara riil melalui matriks kontingensi (*Confusion Matrix*) yang disajikan pada Tabel 2.

Tabel 2. Confusion Matrix

Aktual	Prediksi Gagal(<7.0)	Prediksi Sukses(>7.0)
Gagal	571(TN)	23(FP)
Sukses	99(FN)	60(TP)

C. Pembahasan Hasil Evaluasi Model

Merujuk pada data empiris Tabel 2, tingkat akurasi global sebesar 83,8% mengonfirmasi bahwa karakteristik parameter produksi film yang diolah memiliki korelasi informasi yang sangat kuat untuk memproyeksikan kualitas dan kesuksesan film. Model terbukti sangat kuat dalam mengenali proyek film yang berpotensi masuk ke dalam kategori kurang memuaskan atau gagal (rating di bawah 7,0). Hal ini ditunjukkan oleh tingginya angka *True Negative* (TN) di mana sebanyak 571 data aktual film gagal berhasil didefinisikan secara tepat oleh model, dengan tingkat kesalahan *False Positive* (FP) yang sangat minim, yakni hanya 23 sampel data yang salah diprediksi sebagai film sukses.

Sebaliknya, pada kelas film sukses (rating $\geq 7,0$), model berhasil mengklasifikasikan 60 data secara akurat sebagai *True Positive* (TP). Walau demikian, masih terdapat data *False Negative* (FN) sebanyak 99 film, di mana model salah memproyeksikan film yang secara aktual sukses sebagai film gagal. Secara keseluruhan, kemampuan model yang sangat tajam dalam mendeteksi nilai TN (mencapai 571 sampel) memberikan nilai keandalan yang tinggi bagi produser maupun investor untuk mendeteksi dan memitigasi risiko potensi kegagalan proyek film sejak fase pra-produksi.

4. Analisis Features Importance

Sebagai bagian dari tahapan evaluasi dalam kerangka kerja CRISP-DM, tingkat kontribusi dari setiap variabel produksi dihitung secara matematis menggunakan fitur *feature importance* bawaan dari algoritma *Random Forest Classifier*. Berdasarkan hasil ekstraksi bobot signifikansi pada sistem, atribut durasi film (runtime) atau "Durasi (Menit)" serta tingkat "Popularitas" memiliki peran paling vital dan dominan dengan nilai skor kepentingan tertinggi (mendekati nilai 0,30).

Secara kontekstual, durasi penayangan film yang proporsional memiliki korelasi linear terhadap tingkat kenyamanan penonton di dalam

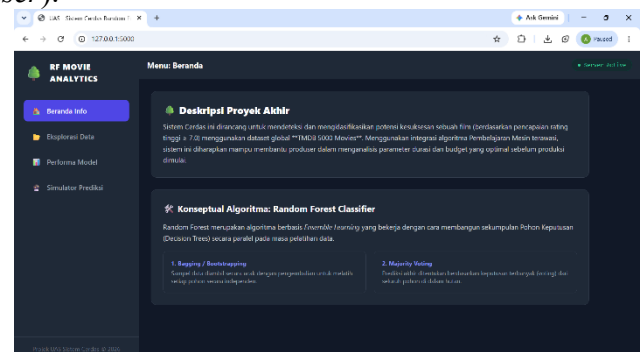
gedung bioskop, yang pada akhirnya memengaruhi ulasan evaluasi rating akhir. Di sisi lain, skor popularitas merepresentasikan efektivitas strategi pemasaran pra-rilis dan tingginya atensi awal masyarakat terhadap proyek film tersebut. Variabel modal awal seperti Budget dan estimasi Revenue berada pada klaster kepentingan tingkat menengah. Menariknya, fitur seperti "Genre Populer" dan keterlibatan "Aktor Kelas Atas" mencatatkan skor kepentingan paling rendah pada model ini. Temuan ini membuktikan secara empiris bahwa alokasi anggaran yang besar untuk merekrut aktor mahal atau berfokus pada genre tren tertentu tidak menjadi jaminan mutlak bagi sebuah film untuk mendapatkan ulasan rating yang tinggi dan masuk ke dalam kategori film sukses.

D. Deployment

Tahap akhir dari siklus standar industri CRISP-DM adalah penyebaran model ke lingkungan pengguna (*Deployment*). Arsitektur model *Random Forest Classifier* dengan tingkat akurasi 83,8% disimpan ke dalam bentuk format berkas biner terserialisasi (.pkl) dengan bantuan pustaka pickle. Model biner tersebut kemudian diintegrasikan ke dalam sisi *backend* aplikasi berbasis web menggunakan *framework* Python Flask untuk mentransformasikan model statistik menjadi sistem pendukung keputusan (*decision support system*) yang interaktif. Implementasi fungsionalitas antarmuka sistem diuraikan menjadi empat komponen utama sebagai berikut:

1. Halaman Beranda

Halaman ini dirancang sebagai menu pembuka utama (*landing page*) yang menyajikan landasan teoretis proyek kepada pengguna akhir (*end-user*).

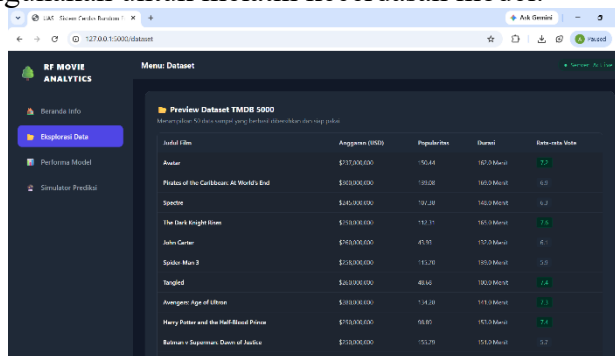


Gambar 3.1. Antarmuka Halaman Beranda Utama Sistem

Gambar 3.1 menampilkan visualisasi konseptual mengenai proyek analisis film. Halaman ini mendefinisikan batasan operasional target kelas sukses ($\text{Rating} \geq 7,0$) dan kelas gagal ($\text{Rating} < 7,0$) menggunakan basis data *TMDB 5000 Movies*. Menu ini juga berfungsi mengedukasi pengguna mengenai mekanisme kerja internal algoritma *Random Forest* melalui pilar *Bagging/Bootstrapping* dan *Majority Voting*.

2. Halaman Eksplorasi Data

Menu ini disediakan untuk memberikan transparansi informasi data historis yang digunakan untuk melatih kecerdasan model.



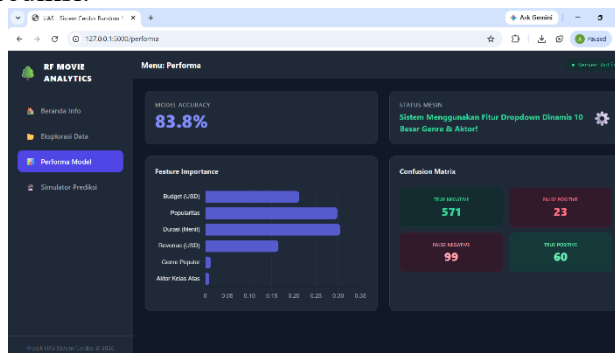
Judul Film	Anggaran (USD)	Populeritas	Durasi	Rata-rata Nilai
Avatar	\$171.000.000	150,44	162,3 Menit	7,9
Princess of the Caribbean: At World's End	\$140.000.000	100,00	140,0 Menit	6,0
Spider-Man	\$140.000.000	132,20	140,0 Menit	6,2
The Dark Knight Rises	\$108.000.000	131,11	165,0 Menit	7,6
John Carter	\$150.000.000	45,93	120,0 Menit	6,1
Sucker Punch	\$100.000.000	11,70	100,0 Menit	5,0
Tangled	\$100.000.000	49,40	100,0 Menit	7,4
Avengers: Age of Ultron	\$150.000.000	114,30	141,0 Menit	7,3
Harry Potter and the Half-Blood Prince	\$150.000.000	68,87	153,0 Menit	7,4
Batman v Superman: Dawn of Justice	\$150.000.000	112,70	153,0 Menit	7,2

Gambar 3.2. Antarmuka Halaman Preview Dataset

Gambar 3.2 menyajikan visualisasi pratinjau (*preview*) dari 50 sampel data historis yang telah bersih setelah melalui tahap *data preparation*. Data disusun dalam bentuk tabel terstruktur yang memuat informasi judul film, nilai anggaran, skor popularitas, durasi, dan rata-rata rating awal yang berkolorasi penuh dengan visualisasi data pada Tabel 1.

3. Halaman Perfoma Model

Halaman ini memuat visualisasi metrik performa guna memberikan jaminan akuntabilitas ilmiah terhadap keandalan sistem prediksi.

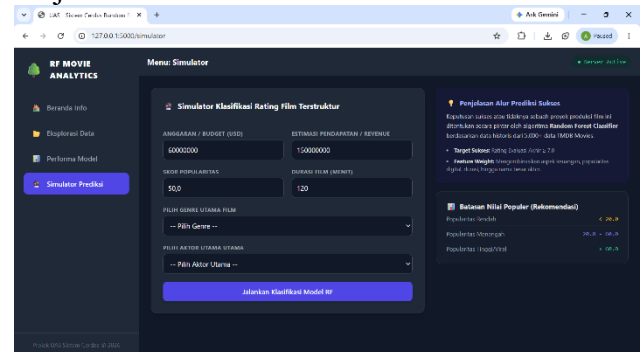


Gambar 3.3. Antarmuka Halaman Dashboard Performa Model

Gambar 3.3 merupakan *dashboard* evaluasi model yang menampilkan tiga informasi kunci: nilai *Model Accuracy* global sebesar **83,8%**, grafik batang horizontal dari nilai *Feature Importance* atribut produksi, serta representasi blok matriks kontingensi (*Confusion Matrix*) hasil pengujian komputer yang sinkron dengan data numerik pada Tabel 2.

4. Halaman Simulator Prediksi

Halaman ini merupakan implementasi fungsionalitas inti dari tahap *deployment* yang memungkinkan pengguna melakukan simulasi manajemen risiko secara mandiri.



Simulator Klasifikasi Rating Film Terstruktur

Anggaran / Budget (USD): 60000000 | Estimasi Pendapatan / Revenue: 150000000

Skor Popularitas: 50,0 | Durasi (Menit): 120

Pilih Genre Utama Film: -- Pilih Genre --

Pilih Aktor Utama Utama: -- Pilih Aktor --

Pilih Aktor Utama Utama: -- Pilih Aktor --

Simulasi Hasil Prediksi Sukses

Target Sukses: Rating Sukses >= 7,0

Realisasi Sukses: 83,8%

Realisasi Gagal: 16,2%

Gambar 3.4. Antarmuka Form Simulator Prediksi

Gambar 3.4 menunjukkan halaman simulator klasifikasi film. Pengguna dapat menginput estimasi parameter produksi (Anggaran, Pendapatan, Skor Popularitas, Durasi) serta memilih rumpun Genre Utama dan Aktor Utama melalui fitur *dropdown* dinamis. Ketika tombol eksekusi ditekan, skrip Flask akan menangkap input secara *real-time*, mengumpulkannya ke model, dan mengembalikan hasil proyeksi klasifikasi (kategori sukses atau gagal) secara instan ke layar pengguna.

IV. KESIMPULAN

Berdasarkan rangkaian eksperimen yang telah dilakukan, penelitian ini berhasil menerapkan metodologi Cross-Industry Standard Process for Data Mining (CRISP-DM) secara utuh sebagai kerangka kerja terstruktur untuk membangun model klasifikasi biner kesuksesan film. Penerapan algoritma Random Forest Classifier dengan konfigurasi parameter $n_estimators=100$

dan `random_state=42` terbukti mampu menghasilkan performa klasifikasi yang sangat baik dan andal. Hal ini ditunjukkan oleh pencapaian nilai akurasi global (Model Accuracy) yang tinggi, yaitu sebesar 83,8% dalam memproyeksikan potensi kesuksesan rating film menggunakan basis data TMDb 5000 Movies.

Melalui analisis kedalaman fitur (Feature Importance), ditemukan secara empiris bahwa atribut durasi film (runtime) dan tingkat popularitas di masyarakat memegang peran paling dominan dalam memengaruhi keputusan percabangan pohon keputusan. Sebaliknya, variabel seperti genre populer dan keterlibatan aktor kelas atas mencatatkan skor kepentingan yang sangat minim. Temuan ini membuktikan secara ilmiah bahwa alokasi anggaran besar untuk merekrut aktor mahal tidak menjadi jaminan mutlak bagi sebuah proyek film untuk mendapatkan evaluasi rating yang tinggi dari audiens. Selain itu, model ini terbukti sangat kuat dan tajam dalam mengenali proyek film yang berpotensi masuk ke dalam kategori gagal (rating di bawah 7,0) melalui pencapaian 571 sampel True Negative (TN). Kemampuan deteksi risiko yang tinggi ini memberikan nilai keandalan yang signifikan bagi para produser maupun investor untuk memitigasi potensi kerugian finansial sejak fase pra-produksi.

Sebagai tahap akhir dari siklus standar industri, proses deployment yang diwujudkan melalui pengintegrasian berkas biner model .pkl ke dalam backend aplikasi web berbasis framework Python Flask telah berhasil menyediakan sistem pendukung keputusan (decision support system) yang interaktif. Sistem ini siap digunakan secara praktis dan real-time oleh pengguna akhir untuk melakukan simulasi karakteristik produksi film. Guna pengembangan penelitian selanjutnya, disarankan untuk menerapkan teknik penanganan

ketidakseimbangan data (imbalanced data) seperti SMOTE (Synthetic Minority Over-sampling Technique). Langkah tersebut diharapkan dapat meningkatkan sensitivitas model dalam mengenali sampel kelas minoritas (film sukses) secara lebih seimbang dan presisi.

REFERENSI

- [1] Y. Gao, "Film as a Form of Cultural Medium : Trends of the Film Industry," 2024.
- [2] Y. Zhang, "Analysis of Financial Risk Avoidance in the Film and TV Industry Based on Big Data," vol. 0, pp. 301–305, 2023.
- [3] D. Ariatmanto and M. Ilham Arief, "Prediksi Peluang Kesuksesan Film Dalam Pra Produksi Menggunakan Algoritma Decision Tree," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 1, pp. 222–226, 2023.
- [4] H. Athaya, I. T. A. Amrulloh, and R. Mufidah, "Prediksi Kesuksesan Film Indonesia Menggunakan Jaringan Syaraf Tiruan," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 2, pp. 2106–2111, 2024.
- [5] F. D. Sianipar *et al.*, "ANALISIS POPULARITAS GENRE FILM DI INDONESIA MENGGUNAKAN ALGORITMA DECISION TREE," vol. 9, no. 4, pp. 5555–5563, 2025.
- [6] N. N. Marpid, Y. I. Kurniawan, and S. P. Rahayu, "ANALYSIS OF THE MOVIE DATABASE FILM RATING PREDICTION WITH ANALISIS PREDIKSI THE MOVIE DATABASE RATING FILM DENGAN MENGGUNAKAN ENSEMBLE LEARNING MENGGUNAKAN METODE RANDOM," vol. 6, no. 1, pp. 1–9, 2025.
- [7] R. N. Rahmah, P. N. Sabrina, and E. Ramadhan, "Prediksi Pendapatan Film Menggunakan Gradient Boosting," *Algoritma*, pp. 1954–1966, 2025, doi: 10.33364/algoritma/v.22-2.2613.
- [8] R. Diyah Ruswanti, Dahlan Susilo, "Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45," *Ilm. STMIK AUB*, vol. 30, no. 1, pp. 111–121, 2024, doi: 10.36309/goi.v30i1.266.
- [9] R. A. Lubis and M. I. P. Nasution, "Penerapan Upaya Pengolahan Kualitas Data Untuk Meningkatkan Informasi yang Konsisten," *Penelit. Ilmu-Ilmu Sos.*, vol. 02, no. June, pp. 207–214, 2025.
- [10] T. Wang, "Improved random forest classification model combined with C5.0 algorithm for vegetation feature analysis in non - agricultural environments," *Sci. Rep.*, pp. 1–13, 2024, doi: 10.1038/s41598-024-60066-x.
- [11] Y. Zheng, "Predicting Movie Box Office Based on Machine Learning , Deep Learning , and Statistical Methods," *Appl. Comput. Eng.*, vol. 0, pp. 20–32, 2024, doi: 10.54254/2755-2721/94/2024MELB0069.
- [12] J. Tripathi, S. Tiwari, A. Saini, and S. Kumari, "Prediction of movie success based on machine learning and twitter sentiment analysis using internet movie database data," vol. 29, no. 3, pp. 1750–1757, 2023, doi: 10.11591/ijeecs.v29.i3.pp1750-1757.
- [13] M. H. Shahid and M. A. Islam, "Investigation of time series-based genre popularity features for box office success prediction," *Comput. Sci.*, vol. 9, 2023, doi: 10.7717/peerj-cs.1603.