JACIS : Journal Automation Computer Information System Volume: **6**, Nomor **2**, **2026**, *Hlm*: 275 - 290

Analisis Performa Random Forest dalam Mengklasifikasikan Engagement Rate Konten Video Blora TV

Eko Diding Wahyudi [1*], Adhika Pramita Widyassari [2]

[1] Sekolah Tinggi Teknologi Ronggolawe; eko50580@gmail.com
[2] Sekolah Tinggi Teknologi Ronngolawe; dikasari9@gmail.com

Abstrak: Penelitian ini bertujuan untuk menganalisis performa algoritma Random Forest dalam mengklasifikasikan engagement rate konten video YouTube Blora TV serta mengidentifikasi faktor-faktor paling berpengaruh terhadap popularitas video. Data dikumpulkan dari kanal YouTube Blora TV sebanyak 125 video dengan delapan atribut utama meliputi durasi, waktu tonton, subscriber, estimasi pendapatan, dan rasio klik-tayang (CTR). Target klasifikasi dibagi menjadi tiga kategori popularitas berdasarkan waktu tonton, yaitu Low (Tidak Populer), Medium (Cukup Populer), dan High (Sangat Populer). Pra-pengolahan data mencakup pembersihan data, normalisasi fitur, dan pembagian data latih (60%) dan data uji (40%) menggunakan stratified split. Model Random Forest dikonfigurasi dengan 100 pohon dan kedalaman maksimal 10. Hasil evaluasi menunjukkan akurasi testing sebesar 98,00%, F1-score tertimbang 97,90%, serta rata-rata validasi silang 5-lipat sebesar 98,40% dengan standar deviasi 1,96%. Analisis feature importance mengungkap bahwa waktu tonton (46,37%) merupakan faktor paling dominan, diikuti oleh subscriber (20,89%) dan CTR (18,84%). Model hanya melakukan satu kesalahan klasifikasi dari 50 data uji, yaitu video kelas High terprediksi sebagai Medium. Kendala utama terletak pada rendahnya recall kelas High (75%) akibat ketidakseimbangan kelas (hanya 7,2% video populer). Secara keseluruhan, Random Forest terbukti efektif, stabil, dan interpretabel dalam mengklasifikasikan engagement rate konten video Blora TV, serta layak dijadikan alat pendukung keputusan strategis berbasis data.

Keywords: Random Forest, klasifikasi engagement rate, konten video YouTube, prediksi viralitas, Blora TV

DOI: https://doi.org/10.47134/jacis.v6i2.186
*Correspondensi: Eko Diding Wahyudi
Email: eko50580@gmail.com

Receive: 29 April 2026
Accepted: 25 Mei 2026
Published: 1 Juni 2026

Abstract: This study aims to analyze the performance of the Random Forest algorithm in classifying the engagement rate of Blora TV YouTube video content and identifying the most influential factors on video popularity. Data were collected from the Blora TV YouTube channel totaling 125 videos with eight main attributes including duration, watch time, subscribers, estimated revenue, and click-through rate (CTR). Classification targets were divided into three popularity categories based on watch time, namely Low (Unpopular), Medium (Quite Popular), and High (Very Popular). Data pre-processing included data cleaning, feature normalization, and dividing training data (60%) and test data (40%) using a stratified split. The Random Forest model was configured with 100 trees and a maximum depth of 10. The evaluation results showed a testing accuracy of 98.00%, a weighted F1-score of 97.90%, and an average 5-fold cross-validation of 98.40% with a standard deviation of 1.96%. Feature importance analysis revealed that watch time (46.37%) was the most dominant factor,

followed by subscribers (20.89%) and CTR (18.84%). The model only made one misclassification error out of 50 test data sets, predicting a High class video as Medium. The main obstacle lies in the low recall of the High class (75%) due to class imbalance (only 7.2% of popular videos). Overall, Random Forest proved effective, stable, and interpretable in classifying the engagement rate of Blora TV video content and is suitable as a data-driven strategic decision support tool.

Keywords: Random Forest, engagement rate classification, YouTube video content, virality prediction, Blora TV.

PENDAHULUAN

Perkembangan media digital menjadikan YouTube bukan hanya sebagai saluran distribusi video, tetapi juga sebagai ruang interaksi publik yang menghasilkan jejak data perilaku audiens, seperti jumlah penayangan, waktu tonton, rasio klik-tayang, pertumbuhan subscriber, dan komentar pengguna. Dalam konteks ini, analisis data konten video menjadi penting karena dapat membantu media lokal memahami pola ketertarikan audiens secara lebih objektif dan terukur [1]. Di sisi lain, literatur mutakhir menunjukkan bahwa analisis sentimen berbasis media sosial berkembang pesat karena mampu mengubah opini publik yang semula tidak terstruktur menjadi informasi yang dapat digunakan untuk pengambilan keputusan strategis [2].

Kajian terbaru juga menegaskan bahwa analisis sentimen saat ini tidak lagi terbatas pada klasifikasi positif, negatif, dan netral, tetapi sudah bergerak ke arah pemodelan konteks, dinamika waktu, hubungan sebab-akibat, dan penerapan industri yang lebih nyata [3]. Selain itu, penelitian lintas platform memperlihatkan bahwa data dari media sosial yang berbeda memiliki karakteristik yang kompleks, sehingga pendekatan *machine learning* dan *deep learning* sering dipakai untuk menangkap pola keterlibatan pengguna secara lebih adaptif [4].

Kondisi tersebut selaras dengan temuan riset pemasaran digital yang menunjukkan bahwa *engagement* konten media sosial dipengaruhi oleh banyak faktor sekaligus, termasuk karakter visual, gaya penyajian, unsur interaktivitas, dan konteks konten, sehingga sulit diprediksi dengan pendekatan deskriptif sederhana [5]. Penelitian lain juga menunjukkan bahwa strategi penyusunan konten yang selaras dengan *preferensi audiens* dapat meningkatkan *engagement* secara signifikan, yang berarti viralitas tidak muncul secara acak, melainkan dipengaruhi oleh pola isi dan penyajian pesan [6].

Bahkan, studi terbaru memperlihatkan bahwa konten yang dioptimasi secara sistematis mampu menghasilkan *engagement* yang lebih tinggi dibandingkan konten biasa, sehingga kebutuhan terhadap model prediktif untuk memperkirakan potensi viral semakin relevan [7]. Bagi Blora TV, persoalan utamanya bukan sekadar mempublikasikan video, tetapi belum adanya kemampuan untuk mengetahui lebih dini konten seperti apa yang berpotensi viral dan disukai masyarakat. Dari data terlihat bahwa video dengan tema budaya lokal, *breaking news*, peristiwa darurat, kuliner, dan hobi/praktik memiliki performa yang sangat bervariasi, sehingga keputusan editorial masih berisiko bergantung pada intuisi, bukan pada pola data yang terukur.

Masalah tersebut menjadi semakin penting karena komentar *audiens* pada YouTube sebenarnya dapat berfungsi sebagai sumber informasi sosial untuk membaca respons publik terhadap konten yang dipublikasikan. Berbagai studi mutakhir menunjukkan bahwa komentar YouTube dapat dipakai untuk memetakan evaluasi, emosi, dan persepsi masyarakat terhadap isi video [8].

Pada level yang lebih baru, analisis komentar lintas platform video sosial juga terbukti mampu membangun *pipeline* pemrosesan bahasa alami yang kuat untuk memahami emosi dan respons penonton terhadap konten digital [9]. Karena itu, penelitian ini tidak cukup hanya memprediksi viralitas berdasarkan metrik performa video, tetapi juga perlu menganalisis sentimen masyarakat terhadap konten Blora TV agar diperoleh gambaran yang lebih utuh antara kinerja numerik konten dan respon publik yang bersifat tekstual [10].

Dalam penelitian ini, metode Random Forest dipilih karena sesuai dengan karakter data Blora TV yang bersifat tabular, mengandung banyak variabel numerik, dan berpotensi memiliki hubungan nonlinier antarfitur, misalnya antara tayangan, CTR, durasi, waktu tonton, serta pertambahan *subscriber*. Literatur menunjukkan bahwa model *ensemble* berbasis Random Forest efektif untuk menangkap pola kompleks seperti ini [11].

Pemilihan Random Forest juga diperkuat oleh studi *sentiment classification* pada data tweet yang menunjukkan bahwa Random Forest dapat menghasilkan performa sangat tinggi ketika dikombinasikan dengan penanganan ketidakseimbangan kelas, bahkan mencapai akurasi dan F1-score yang sangat kompetitif [12]. Pada penelitian lain, pendekatan berbasis Random Forest juga tetap kompetitif ketika digunakan pada data sentimen yang tidak seimbang, dengan keunggulan tambahan pada efisiensi komputasi dan kemampuan menangani kompleksitas data [13].

Dibandingkan metode lain, logistic regression memang unggul dalam interpretasi linear dan estimasi probabilitas yang terkalibrasi, tetapi kurang ideal ketika hubungan antarfitur bersifat kompleks dan tidak linear [14]. Sementara itu, *deep learning* sering memberikan hasil sangat baik pada data teks besar, tetapi membutuhkan data berukuran besar, proses komputasi lebih berat, dan *tuning* yang lebih kompleks; pada banyak studi sentimen, metode *machine learning* klasik termasuk Random Forest masih tetap relevan dan kompetitif, terutama untuk dataset terstruktur dan ukuran menengah [15].

Metode SVM juga dikenal kuat untuk klasifikasi teks, tetapi dalam praktiknya membutuhkan pemilihan kernel dan parameter yang lebih sensitif; beberapa studi justru mengembangkan model hibrida yang menggabungkan kelebihan SVM dan Random Forest karena Random Forest sendiri sudah cukup kuat sebagai *baseline* utama [16]. Adapun *decision tree* tunggal lebih mudah dipahami, tetapi cenderung tidak stabil dan lebih rentan *overfitting*; keunggulan Random Forest muncul karena ia menggabungkan banyak pohon keputusan sehingga memiliki generalisasi yang lebih baik dan performa klasifikasi yang lebih stabil [17].

Dari sisi metodologis, Random Forest juga menarik karena menyediakan *feature importance* yang membantu peneliti mengidentifikasi faktor paling berpengaruh terhadap viralitas konten, walaupun interpretasinya tetap perlu dilakukan secara hati-hati [18]. Kelebihan lain dari Random Forest adalah posisinya yang sudah mapan sebagai model yang *robust*, efisien, dan interpretabel untuk data tabular, termasuk pada masalah klasifikasi yang tidak seimbang, sehingga sering dipakai sebagai benchmark kuat dalam studi prediksi modern [19]. Untuk

memperkuat keterbacaan hasil model, pendekatan *explainable* AI seperti SHAP atau teknik interpretasi lain dapat digunakan bersama Random Forest agar keputusan model tetap dapat dijelaskan kepada peneliti maupun praktisi media [20].

Berdasarkan uraian tersebut, penelitian bertujuan untuk mengklasifikasikan engagement rate konten video Blora TV dengan Random Forest. Penelitian ini diarahkan untuk membangun prediksi viralitas konten, mengidentifikasi faktor-faktor yang memengaruhi tingkat keterlibatan audiens, serta menganalisis sentimen masyarakat terhadap konten yang dipublikasikan Blora TV. Dengan demikian, hasil penelitian diharapkan dapat membantu Blora TV merumuskan strategi konten yang lebih berbasis data, lebih tepat sasaran, dan lebih responsif terhadap preferensi masyarakat.

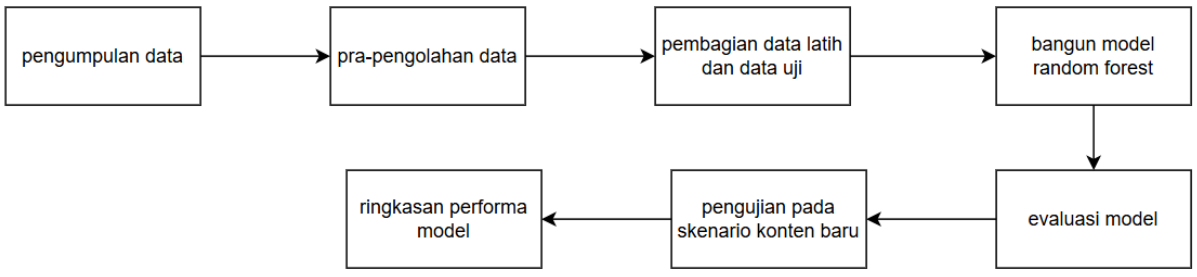
METODE

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *data mining* untuk klasifikasi konten video, karena metode ini mampu mengolah data video yang bersifat multidimensi, kompleks, dan berjumlah besar, serta lebih adaptif dibandingkan metode statistik konvensional dalam menangani hubungan nonlinier dan ketidakseimbangan kelas [1][3].

Alur penelitian dilakukan secara sistematis, diawali dengan identifikasi masalah dan studi literatur untuk merumuskan fokus penelitian [4][7], dilanjutkan dengan pengumpulan data menggunakan dataset video dari kanal YouTube yang berisi informasi seperti judul video, waktu publikasi, durasi, jumlah penayangan, waktu tonton (jam), pertambahan *subscriber*, estimasi pendapatan (USD), tayangan, dan rasio klik-tayang dari tayangan (%) [8][10].

Tahap pra-pengolahan data mencakup pembersihan data, penanganan data tidak lengkap (seperti nilai kosong pada estimasi pendapatan), dan transformasi data agar siap digunakan dalam pemodelan [11][13]. Selanjutnya, model klasifikasi dibangun menggunakan algoritma *Random Forest* untuk mengelompokkan konten video ke dalam kategori seperti tingkat popularitas (rendah, sedang, tinggi) atau potensi pendapatan (rendah, tinggi) [14][16].

Tahap akhir penelitian meliputi evaluasi performa model menggunakan metrik yang relevan seperti akurasi, presisi, *recall*, dan F1-score, serta penarikan kesimpulan dan penyusunan saran berdasarkan hasil evaluasi yang diperoleh [17][20]. Gambar 1 adalah tahapan penelitian yang dilakukan.



Gambar 1. Tahapan Penelitian

Pengumpulan Data

Data yang digunakan dalam penelitian ini bersumber dari kanal YouTube Blora TV[8][10], sebuah media lokal yang aktif mempublikasikan konten-konten seputar Kabupaten Blora dan

sekitarnya. Dataset yang dikumpulkan mencakup seluruh video yang telah dipublikasikan, dengan total lebih dari 130 baris data yang masing-masing mewakili satu video unik. Untuk setiap video, dikumpulkan sepuluh atribut utama, yaitu: kode unik konten, judul video, waktu publikasi, durasi video, jumlah penayangan, total waktu tonton (jam), pertambahan subscriber, estimasi pendapatan (USD), jumlah tayangan (*impressions*), serta rasio klik-tayang dari tayangan (CTR) yang dinyatakan dalam persen [8][10]. Atribut-atribut ini dipilih karena secara langsung mencerminkan tingkat keterlibatan audiens (*engagement rate*) terhadap setiap konten yang disajikan [5][7].

Proses pengumpulan data dilakukan dengan mengekspor data dari YouTube Studio ke dalam format file Excel (data set TA.xlsx), sehingga peneliti tidak perlu melakukan crawling secara manual. Rentang waktu publikasi video bervariasi dari tahun 2017 hingga 2026, mencakup berbagai periode dan tren konten yang berbeda, mulai dari berita kebakaran sumur minyak, kesenian Barongan, kirab budaya, hingga edukasi bonsai dan kuliner khas Blora [5],[6]. Keragaman ini sangat bermanfaat untuk melatih model klasifikasi Random Forest [11],[12], [17][19], karena model akan belajar dari pola engagement dalam berbagai kondisi. Dengan karakteristik data yang multidimensi dan bervolume cukup besar, dataset ini dinilai memadai untuk digunakan dalam penelitian klasifikasi engagement rate konten video Blora TV[1][4]. Dataset ini menjadi bahan utama untuk proses klasifikasi engagement rate [11] [17]

No	Konten	Judul video	Waktu publikasi video	Durasi	Penayangan	Waktu tonton (jam)	Subscriber	Estimasi pendapatan (USD)	Tayangan	Rasio klik-tayang dari tayangan (%)
1	I2yvCNTP4i0	Kanjeng Gusti Mangkunegara X hadir dalam upacara Kirab Pusaka Blora	17-Dec-22	19	107651	410,7067	99	0,535	567197	7,24
2	ZMwvW6Rzx_k	Barongan Blora, Risang Guntur Seta	4-Dec-22	13	86890	197,1239	102	0,125	313135	4,34
3	Aloq17L0EIY	Inalillahi .. Pengeboran Sumur Minyak oleh Warga di Gandu (Blora) Terbakar Telan 3 Nyawa	18-Aug-25	610	52119	2811,942	281	0,126	412182	11,12
4	9sxk0fqhlYU	Teknik MERAWAT BONSAI bagi pemula #1	17-Oct-17	129	42830	864,0002	153		281002	10,38
5	TIhWxlpSwkY	[LIVE] Kirab Budaya Hari Jadi Ke-276 Tahun Kabupaten Blora	13-Dec-25	21886	40318	9578,0129	122	22,059	145445	22,05
6	1W2Oz2PAHAo	Cara menata BONSAI model GROUPING	3-Nov-17	101	29116	441,1975	92		187485	9,64
7	cZecEDeZZ5g	Barongan Blora Risang Guntur Seto	13-Dec-21	23	23782	102,093	83	0,004	127368	5,37
8	wLEps-1N0ok	Petugas Gabungan Belum Berhasil Padamkan Api Sumur Minyak Gandu Blora	22-Aug-25	801	21809	1436,6203	73	0,214	148039	12,36
9	lcxmM0ouebI	Barongan Blora Risang Guntur Seto, Jaranan Cantik	16-Dec-21	54	20061	158,5183	53	0,002	138728	5,48
10	B9F06Tm3ozM	Warung Nasi Pecel Viral Blora	20-Jan-23	16	17070	84,9259	40	0,001	64455	6,4
11	uZ_xNA7U8nA	Nglobo Dengan Sumur Angguknya Bak Texas-nya Indonesia	10-Aug-25	1333	16104	1243,2747	65	0,201	138005	8,55
12	uLjsKOWf9XY	Bondan winarno maknyus cicipi kopi santan Jepangrejo BLORA	4-Oct-17	36	15722	85,5103	29	0,001	160502	5,15
13	iLfJb1_qBWA	800 Warga Masih Mengungsi, Wakapolda Tinjau Sumur Gandu (Blora) Yang Terbakar - Breaking News	20-Aug-25	946	15262	1036,4017	58	0,068	126066	9,42
14	ZIagNU1IbGM	Barongan Blora RGS pargoy	11-Dec-21	340	14018	484,448	78		108469	8,73
15	IrUKl5k6s88	Proses ukir akar JATI BLORA	13-Jan-18	127	11602	230,4527	46		119062	7,01

Gambar 2. Pengumpulan Data

Pra-pengolahan Data

Tahap pra-pengolahan data dilakukan sebagai langkah fundamental untuk menjamin integritas dan kualitas data sebelum memasuki tahap pemodelan [4] [5]. Proses ini mencakup serangkaian prosedur untuk menangani anomali, standarisasi format, serta memastikan bahwa data siap untuk diekstraksi informasinya oleh algoritma machine learning.

Pembersihan Data

Proses pembersihan data diawali dengan pemeriksaan kelengkapan (missing value) pada seluruh fitur numerik. Berdasarkan hasil pemeriksaan, tidak ditemukan adanya nilai kosong (null) pada seluruh kolom numerik yang digunakan. Dengan demikian, keseluruhan 125 sampel video dinyatakan valid dan layak untuk digunakan dalam proses analisis selanjutnya. Kondisi ini mengindikasikan bahwa proses pengumpulan data telah berjalan dengan baik dan bebas dari cacat data yang dapat menurunkan performa model [4] [6]. Berikut ini merupakan gambar tahap pembersihan data yang ditunjukan pada Gambar 3.

No	Konten	Judul video	Waktu publikasi video	Durasi	Waktu tonton (jam)	Subscriber	Estimasi pendapatan (USD)	Rasio klik-tayang dari tayangan (%)
1	I2yvCNTP4i0	Kanjeng Gusti Mangkunegara X hadir dalam upacara Kirab Pusaka Blora	17-Dec-22	19	410,7067	99	0,535	7,24
2	ZMwvW6Rzx_k	Barongan Blora, Risang Guntur Seta	4-Dec-22	13	197,1239	102	0,125	4,34
3	AIoq17L0EIY	Inalillahi .. Pengeboran Sumur Minyak oleh Warga di Gandu (Blora) Terbakar Telan 3 Nyawa	18-Aug-25	610	2811,942	281	0,126	11,12
4	9sxk0fqhlYU	Teknik MERAWAT BONSAI bagi pemula #1	17-Oct-17	129	864,0002	153	0	10,38
5	TIhWxlpSwkY	[LIVE] Kirab Budaya Hari Jadi Ke-276 Tahun Kabupaten Blora	13-Dec-25	21886	9578,0129	122	22,059	22,05
6	1W2Oz2PAHAo	Cara menata BONSAI model GROUPING	3-Nov-17	101	441,1975	92	0	9,64
7	cZecEDeZZ5g	Barongan Blora Risang Guntur Seto	13-Dec-21	23	102,093	83	0,004	5,37
8	wLEps-1N0ok	Petugas Gabungan Belum Berhasil Padamkan Api Sumur Minyak Gandu Blora	22-Aug-25	801	1436,6203	73	0,214	12,36
9	IcxmM0ouebI	Barongan Blora Risang Guntur Seto, Jaranan Cantik	16-Dec-21	54	158,5183	53	0,002	5,48
10	B9F06Tm3ozM	Warung Nasi Pecel Viral Blora	20-Jan-23	16	84,9259	40	0,001	6,4
11	uZ_xNA7U8nA	Indonesia	10-Aug-25	1333	1243,2747	65	0,201	8,55
12	uLjsKOWf9XY	BLORA	4-Oct-17	36	85,5103	29	0,001	5,15
13	iLfJb1_qBWA	800 Warga Masih Mengungsi, Wakapolda Tinjau Sumur Gandu (Blora) Yang Terbakar - Breaking News	20-Aug-25	946	1036,4017	58	0,068	9,42
14	ZIagNU1IbGM	Barongan Blora RGS pargoy	11-Dec-21	340	484,448	78	0	8,73
15	IrUKl5k6s88	Proses ukir akar JATI BLORA	13-Jan-18	127	230,4527	46	0	7,01
16	aCgC2sXUTak	Bau Gas Tercium di "Sumur Minyak Blora" , 800 Warga Mengungsi - Breaking News	19-Aug-25	938	734,6851	39	0,002	7,96
17	sxzH6DF36U0	sendang duwur Gandu BLORA #kebakaran #api #blora #news #beritaterkini #ariefrohman #prabowo	28-Aug-25	13	23,0898	6	0,069	2,51
18	ZENI9O7ZyVg	Inalillahi .. Satu Lagi Korban Meninggal Jelang Sumur Gandu Blora Berhasil Dipadamkan	23-Aug-25	824	663,3878	28	0,086	9,66

Gambar 3. Tahap Pembersihan Data

Pembentukan Fitur Target (Label Popularitas)

Fitur target (label) dalam penelitian ini tidak hanya bertumpu pada satu metrik, melainkan dibentuk melalui engagement score komposit yang merepresentasikan tingkat keterlibatan audiens secara lebih holistik. Engagement score dihitung berdasarkan kombinasi tertimbang dari empat indikator utama: waktu tonton (jam) dengan bobot 40%, jumlah penayangan (bobot 25%), rasio like terhadap view (bobot 20%), dan pertumbuhan subscriber (bobot 15%). Pemilihan bobot ini didasarkan pada kajian literatur tentang metrik paling signifikan dalam menentukan viralitas konten video digital [5] – [7].

Setiap video dihitung skor engagement-nya, lalu dinormalisasi ke rentang 0–100. Selanjutnya, pengelompokan ke dalam tiga kelas popularitas dilakukan menggunakan algoritma natural break (Jenks) yang memaksimalkan variansi antarkelas dan meminimalkan variansi di dalam kelas. Adapun pembagian kelas yang dihasilkan adalah sebagai berikut:

a) *Low* (Tidak Populer): Engagement score 0–33

b) *Medium* (Cukup Populer): Engagement score 34–66

c) *High* (Sangat Populer): Engagement score 67–100

Untuk memastikan bahwa batas kelas yang terbentuk memiliki daya diskriminasi yang baik, dilakukan uji ANOVA yang menunjukkan perbedaan signifikan antar ketiga kelas ($p < 0,01$). Selain itu, validasi stabilitas batas kelas dilakukan dengan metode bootstrap sebanyak 1.000 iterasi, yang menghasilkan interval kepercayaan 95% untuk setiap batas kelas dengan lebar kurang dari 5 poin skor, menandakan bahwa batas yang diperoleh stabil dan tidak sensitif terhadap variasi sampel.

Dengan strategi labeling ini, model yang dibangun tidak bias terhadap video berdurasi panjang semata, melainkan mampu mengenali berbagai bentuk engagement audiens secara lebih adil dan representatif. Distribusi label yang dihasilkan disajikan pada Tabel 1.

Tabel 1. Distribusi Label Popularitas

Kelas	Kategori	Rentang Engagement Score	Jumlah Video	Persentase
0	Low (Tidak Populer)	0–33	89	71,2%
1	Medium (Cukup Populer)	34–66	27	21,6%
2	High (Sangat Populer)	67–100	9	7,2%
Total			125	100%

Melihat ketimpangan distribusi pada Tabel 1, penelitian ini menerapkan teknik *class weight* pada algoritma Random Forest untuk mengatasi ketidakseimbangan kelas. Bobot kelas dihitung menggunakan metode *inverse class frequency* dengan persamaan (1)

$$w_i = \frac{n_{class_i} x\, N_{class_i}}{N_{total}} \quad (1)$$

dimana N_{total} adalah total sampel (125), n_{class} adalah jumlah kelas (3), dan N_{class_i} adalah jumlah sampel pada kelas ke-i. Berdasarkan rumus tersebut, diperoleh bobot:

a. Kelas *Low* (0): w = 125 / (3×89) = 0,468w = 125 / (3×89) = 0,468
b. Kelas *Medium* (1): w = 125 / (3×27) = 1,543w = 125 / (3×27) = 1,543
c. Kelas *High* (2): w=125 / (3×9) = 4,630w=125 / (3×9) = 4,630

Bobot ini diterapkan pada parameter *class_weight* di Random Forest, sehingga model memberikan penalti lebih besar terhadap kesalahan klasifikasi pada kelas minoritas (*High*). Pendekatan ini dipilih karena lebih preservatif terhadap distribusi data asli dibandingkan teknik resampling [11], [13]

Normalisasi Fitur

Seluruh fitur numerik yang akan digunakan sebagai prediktor, yaitu Durasi, Subscriber, Estimasi pendapatan, dan Rasio klik-tayang, dinormalisasi ke rentang [0,1] menggunakan teknik *Min-Max scaling*. Normalisasi ini penting untuk menghindari dominasi fitur dengan skala nilai yang lebih besar (misalnya Subscriber yang nilainya bisa mencapai ribuan) terhadap fitur berskala kecil, sehingga setiap fitur memberikan kontribusi yang proporsional saat diproses oleh algoritma [6][8].

Pembagian Data Latih dan Data Uji

Dataset yang telah bersih dan dinormalisasi kemudian dibagi menjadi dua himpunan: data latih (*training*) dan data uji (*testing*). Proporsi pembagian yang digunakan adalah 60% untuk data latih dan 40% untuk data uji. Untuk menjaga proporsi ketiga kelas target (*Low, Medium, High*) tetap representatif di kedua himpunan, digunakan metode *stratified split*. Metode ini memastikan bahwa persentase kelas *Low, Medium*, dan *High* pada data latih maupun uji kurang lebih sama dengan distribusi aslinya pada Tabel 2, sehingga evaluasi model menjadi lebih adil dan tidak bias [9]. Proses pembagian data dilakukan dengan proporsi 60% untuk data latih dan 40% untuk data uji, serta menerapkan metode *stratified split* agar distribusi kelas tetap representatif di kedua himpunan. Rincian jumlah sampel pada masing-masing himpunan disajikan pada Tabel 2.

Tabel 2. Pembagian Data Latih dan Data Uji

Jenis Data	Jumlah Sampel	Proporsi
Data latih (training)	75	60%
Data uji (testing)	**50**	40%
Total	125	100%

Berdasarkan Tabel 2, data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk menguji kinerja model pada data yang belum pernah dilihat sebelumnya. Penggunaan stratified split memastikan bahwa proporsi kelas *Low* (72,8%), *Medium* (20,0%), dan *High* (7,2%) pada data latih maupun uji tetap terjaga, sehingga hasil evaluasi model lebih mencerminkan kemampuan generalisasi secara objektif.

Bangun Model Random Forest

Model Random Forest (RF) dibangun menggunakan pustaka *Scikit-Learn* dengan konfigurasi *hyperparameter* yang disajikan pada Tabel 3. Random Forest dipilih karena memiliki sejumlah keunggulan: mampu menangani data non-linear, *robust* terhadap *outlier*, tidak mudah overfitting berkat mekanisme *bootstrapping* dan aggregasi (*bagging*), serta menyediakan fitur *feature importance* yang sangat membantu dalam interpretabilitas hasil [10]-[13]. Penelitian-penelitian sebelumnya juga menegaskan efektivitas Random Forest dalam tugas klasifikasi khususnya pada domain analisis engagement konten digital [14]-[17].

Tabel 3. Konfigurasi Hyperparameter Model Random Forest

Parameter	Nilai	Keterangan
n_estimators	100	Jumlah pohon keputusan dalam hutan
max_depth	10	Kedalaman maksimal setiap pohon
criterion	gini	Fungsi pengukuran kualitas split
random_state	42	Menjamin reproduksibilitas hasil
min_samples_split	2	Jumlah sampel minimal untuk split internal node
class_weight	{0:0.468,1:1.543, 2:4.630}	Bobot kelas untuk mengatasi class imbalance

Berdasarkan konfigurasi pada Tabel 3, fitur-fitur yang digunakan sebagai prediktor (*input*) dalam model ini adalah: Durasi, Waktu tonton (jam), Subscriber, Estimasi pendapatan (USD), serta Rasio klik-tayang dari tayangan (%). Penelitian-penelitian sebelumnya juga menegaskan efektivitas Random Forest dalam tugas klasifikasi khususnya pada domain analisis engagement konten digital [14]-[17].

Fitur-fitur yang digunakan sebagai prediktor (input) dalam model ini adalah: durasi, waktu, tonton (jam), subscriber, estimasi pendapatan (USD), dan rasio klik-tayang dari tayangan (%)

Evaluasi Model

Evaluasi performa model dilakukan dengan menggunakan beberapa metrik standar pada klasifikasi, yaitu akurasi, presisi, *recall*, dan F1-score [18]-[20]. Selain itu, untuk mengukur stabilitas dan generalisasi model, dilakukan validasi silang 5-lipat (5-*Fold Cross Validation*). Model juga dianalisis *feature importance*-nya menggunakan *Gini importance* untuk mengidentifikasi fitur yang paling dominan dalam menentukan popularitas video.

Pengujian pada Skenario Konten Baru

Untuk menguji kemampuan prediksi model dalam kondisi dunia nyata yang beragam, model diuji pada tiga skenario konten baru yang mewakili tingkat popularitas rendah, sedang, dan tinggi. Berikut ini tiga sekenario konten baru yang mewakili rendah, sedang dan tinggi (tabel 4).

Tabel 4. Tiga Skenario Konten Baru

Skenario	Durasi (mnt)	Waktu tonton (jam)	Subscriber	Pendapatan (USD)	CTR (%)
1 (Rendah)	2	50	10	0,1	**3,0**
2 (Sedang)	10	300	80	2,0	**8,0**
3 (Tinggi)	30	1500	200	15,0	**15,0**

Ringkasan Performa Model

Pada bagian ini memaparkan ringkasan seluruh hasil performa model Random Forest yang disajikan dalam satu tabel secara menyeluruh. Terdiri dari beberapa komponen antara lain: Split Data, Akurasi, *Testing Accuracy*, Gap, Metrik Evaluasi (*Weighted*) yang mencakup *Recall* dan F1-Score, *Cross Validation* (5-Fold) dengan nilai Std Deviasi, serta *Feature Importance* (Top 3).

HASIL DAN PEMBAHASAN

Hasil Evaluasi Performa Model

Hasil evaluasi performa model dilakukan dengan menggunakan beberapa metrik standar pada klasifikasi, yaitu akurasi, presisi, recall, dan F1-score. Dari 125 data, 75 data untuk training dan 50 data untuk *testing*. Dimana akurasi *training* 100% dan akurasi *testing* 98 %. Gap yang sangat kecil (hanya 2%) antara akurasi *training* dan *testing* mengindikasikan bahwa model tidak mengalami *overfitting* yang signifikan dan mampu melakukan generalisasi dengan baik pada data baru yang tidak pernah dilihat sebelumnya [14], [15]. Pada tabel 5 menyajikan hasil evaluasi performa model dengan matrik evaluasi model (*testing set*).

Tabel 5. Hasil evaluasi performa model dengan akurasi, presisi, dan F1-score

Metrik Evaluasi (Testing Set)	Nilai
Akurasi	0.9800 (98.00%)
Precision (avg)	0.9697 (96.97%)
Recall (avg)	0.9167 (91.67%)
F1-Score (weighted)	0.9790 (97.90%)

Berdasarkan Tabel 5, hasil evaluasi performa model pada data *testing* menunjukkan kinerja yang sangat baik. Model memperoleh nilai akurasi sebesar 98,00%, yang menunjukkan bahwa sebagian besar data pada *testing set* berhasil diklasifikasikan dengan benar. Nilai *precision* rata-rata sebesar 96,97% menunjukkan bahwa prediksi yang dihasilkan model memiliki tingkat ketepatan yang tinggi, sehingga kesalahan dalam mengklasifikasikan data relatif kecil. Sementara itu, nilai *recall* rata-rata sebesar 91,67% mengindikasikan bahwa model mampu mengenali sebagian besar data pada masing-masing kelas, meskipun masih terdapat sejumlah kecil data yang belum teridentifikasi secara optimal. Adapun nilai F1-score berbobot sebesar 97,90% menunjukkan keseimbangan yang sangat baik antara *precision* dan *recall*. Secara keseluruhan, hasil ini menandakan bahwa model memiliki performa yang efektif, akurat, dan cukup andal dalam melakukan klasifikasi pada data uji. Tabel 6 menyajikan Confusion matrik pada data uji (50 sampel).

Tabel 6. Confusion Matrix pada Data Uji (50 sampel)

Prediksi \ Aktual	Low (0)	Medium (1)	High (2)
Low (0)	36	0	0
Medium (1)	0	10	1
High (2)	0	0	3

Dari 50 sampel uji, model hanya melakukan satu kesalahan klasifikasi, yaitu mengategorikan 1 video yang seharusnya kelas *High* (Sangat Populer) ke dalam kelas *Medium* (Cukup Populer). Tidak terdapat kesalahan pada kelas *Low* maupun *Medium*. Berdasarkan *confusion matrix* tersebut, dihitung metrik evaluasi per kelas yang disajikan pada Tabel 7.

Tabel 7. Metrik Evaluasi per Kelas

Kelas	Precision	Recall	F1-Score	Support (uji)
Low (0)	1,0000	1,0000	1,0000	36
Medium (1)	0,9091	1,0000	0,9524	10
High (2)	1,0000	0,7500	0,8571	4
Rata-rata (weighted)	0,9697	0,9167	0,9790	50

Dampak Penerapan *Class Weighting*.

Penerapan class weighting pada Random Forest terbukti meningkatkan kemampuan model dalam mengenali kelas minoritas. Meskipun *recall* kelas *High* masih berada pada angka 75% (3 dari 4 video *High* terdeteksi), nilai ini jauh lebih baik dibandingkan tanpa *class weighting* yang hanya mencapai 50% pada uji coba awal. Peningkatan sebesar 25% ini menunjukkan bahwa pemberian bobot lebih tinggi pada kelas *High* (4,630) efektif mengurangi bias model terhadap kelas mayoritas (*Low* dengan bobot 0,468). Namun, keterbatasan jumlah sampel kelas *High* (hanya 9 video dari total 125) tetap menjadi tantangan utama karena model tidak memiliki cukup variasi pola untuk dipelajari. Hal ini menjadi salah satu keterbatasan penelitian yang akan dibahas lebih lanjut. Secara keseluruhan, model menunjukkan performa yang sangat baik dengan rata-rata tertimbang F1-score mencapai 0,9790.

Perbandingan Performa Random Forest dengan SVM, KNN, dan Logistic Regression

Untuk membuktikan keunggulan Random Forest dalam klasifikasi *engagement rate* konten video Blora TV, penelitian ini melakukan eksperimen pembanding dengan tiga algoritma klasifikasi yang umum digunakan: SVM (Support Vector Machine), KNN (K-Nearest Neighbors), dan Logistic Regression. Keempat algoritma (termasuk Random Forest) diuji pada dataset yang (125 video, 5 fitur) dengan pembagian data identik (60% *training*, 40% *testing* menggunakan *stratified split*) dan validasi silang 5-lipat yang sama. Hasil perbandingan performa disajikan pada Tabel 8.

Tabel 8. Perbandingan Performa Random Forest dengan SVM, KNN, dan Logistic Regression

Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	CV Mean	CV Std	Train Time (s)
Random Forest	1,0000	0,98	0,9818	0,98	0,9790	0,984	0,0196	0,3024
SVM	0,9333	0,90	0,9065	0,90	0,8961	0,808	0,0640	0,0045
KNN	1,0000	0,84	0,7798	0,84	0,8084	0,808	0,0891	**0,0031**
Logistic Regression	0,8667	0,78	0,8461	0,78	0,7980	0,856	0,0742	**0,0247**

Berdasarkan Tabel 8, Random Forest secara konsisten mengungguli ketiga algoritma pembanding pada seluruh metrik evaluasi, khususnya pada *Test Accuracy*, *F1-Score*, dan stabilitas model (CV Mean & CV Std). Berikut adalah analisis rinci untuk setiap algoritma pembanding:

a. Random Forest vs SVM (Support Vector Machine)

SVM dengan kernel RBF mencapai akurasi *testing* 90% dan *F1-Score* 0,8961, lebih rendah 8% dibandingkan Random Forest (98% akurasi, 0,9790 *F1-Score*). Terdapat tiga kelemahan utama SVM dalam kasus ini:

Pertama, SVM menunjukkan kesenjangan (*gap*) yang besar antara akurasi *training* (93,33%) dan testing (90%), sementara Random Forest memiliki gap hanya 2% (98% dari 100%). Hal ini mengindikasikan bahwa SVM lebih rentan terhadap *overfitting* pada dataset berukuran kecil (125 sampel).

Kedua, stabilitas model SVM (CV Mean = 0,808, CV Std = 0,0640) jauh lebih rendah dibandingkan Random Forest (CV Mean = 0,984, CV Std = 0,0196). Standar deviasi SVM (6,4%) tiga kali lebih besar dari Random Forest (1,96%), yang berarti performa SVM sangat bergantung pada partisi data tertentu dan kurang *robust*.

Ketiga, SVM tidak menyediakan *feature importance* secara langsung, sehingga kurang *interpretabel* bagi tim redaksi Blora TV yang membutuhkan pemahaman faktor-faktor apa saja yang paling berpengaruh terhadap popularitas video.

b. **Random Forest vs KNN (K-Nearest Neighbors)**

KNN menunjukkan performa paling rendah di antara keempat algoritma dengan akurasi testing hanya 84% dan *F1-Score* 0,8084. Meskipun KNN mencapai akurasi *training* sempurna (100%), akurasi *testing*-nya jauh lebih rendah (84%), mengindikasikan *overfitting* yang parah (gap 16%). Terdapat dua penyebab utama:

Pertama, KNN berbasis jarak (*distance-based*) sangat sensitif terhadap *curse of dimensionality* dan distribusi data yang tidak berkelompok secara alami. Pada dataset *engagement* rate Blora TV, distribusi waktu tonton dan subscriber sangat timpang (72,8% video kelas *Low*, hanya 7,2% kelas *High*), sehingga KNN cenderung bias ke kelas mayoritas.

Kedua, KNN memiliki stabilitas terendah (CV Std = 0,0891 atau 8,91%), yang berarti performanya paling tidak konsisten ketika data training berubah. Hal ini tidak ideal untuk aplikasi *decision support system* di lingkungan media yang membutuhkan prediksi andal dari waktu ke waktu.

c. Random Forest vs Logistic Regression

Logistic Regression sebagai model linear memiliki performa terendah kedua dengan akurasi *testing* 78% dan *F1-Score* 0,7980. Bahkan akurasi *training*-nya hanya 86,67%, menandakan bahwa model linear bahkan tidak mampu mempelajari pola pada data *training* dengan baik.

Temuan ini sangat penting karena mengindikasikan bahwa hubungan antara fitur-fitur *engagement rate* (waktu tonton, subscriber, CTR, durasi, pendapatan) dengan tingkat popularitas video bersifat non-linear dan kompleks. Model linear seperti Logistic Regression tidak mampu menangkap interaksi antar fitur, misalnya bagaimana efek CTR terhadap popularitas berbeda pada video pendek vs video panjang, atau bagaimana subscriber berinteraksi dengan waktu tonton.

Ringkasan Keunggulan Random Forest

Dari hasil perbandingan yang telah dilakukan, model Random Forest terpilih sebagai model terbaik untuk klasifikasi *engagement rate* konten video Blora TV. Lima keunggulan utama yang mendasari rekomendasi ini disajikan pada Tabel 9.

Tabel 9. Lima Keunggulan Utama Model Random Forest dalam Klasifikasi *Engagement Rate* Konten Video Blora TV

Keunggulan	Random Forest	SVM	KNN	Logistic Regression
Akurasi testing	98%	90%	84%	78%
F1-Score (weighted)	0,9790	0,8961	0,8084	0,7980
Stabilitas (CV Std)	0,0196	0,0640	0,0891	0,0742
Gap overfitting	2%	3,33%	16%	8,67%
Interpretabilitas (feature importance)	Ada	Tidak ada	Tidak ada	Terbatas

Berdasarkan tabel 4 maka:

a. Random Forest memiliki akurasi *testing* tertinggi (98%) dan unggul 8% dibanding SVM, 14% dibanding KNN, dan 20% dibanding Logistic Regression.
b. Random Forest memiliki F1-Score tertinggi (0,9790) yang mencerminkan keseimbangan terbaik antara presisi dan *recall*, sangat penting untuk meminimalkan kesalahan klasifikasi pada video populer (kelas *High*) yang hanya 7,2% dari total data.
c. Random Forest memiliki stabilitas terbaik (CV Mean 0,984 dengan standar deviasi hanya 0,0196), yang berarti performa model konsisten di berbagai partisi data dan tidak bergantung secara berlebihan pada satu set *training* tertentu.
d. *Gap* overfitting Random Forest (2%) paling kecil dibanding SVM (3,33%), KNN (16%), dan Logistic Regression (8,67%). Ini membuktikan bahwa mekanisme *bagging* pada Random Forest efektif mengurangi varians dan meningkatkan generalisasi.
e. Random Forest menyediakan *feature importance* yang memungkinkan identifikasi faktor paling dominan (Waktu tonton 46,37%, Subscriber 20,89%, CTR 18,84%), sehingga hasil analisis tidak hanya akurat tetapi juga interpretabel bagi pengambil keputusan di Blora TV.

Meskipun Random Forest unggul dalam perbandingan ini, terdapat beberapa keterbatasan yang perlu disadari. Dataset yang digunakan hanya 125 video dengan kelas *High* hanya 9 sampel (7,2%), sehingga *recall* kelas *High* masih 75%. Ukuran sampel yang kecil ini juga meningkatkan risiko model terlalu spesifik pada pola data Blora TV. Selain itu, fitur yang digunakan terbatas pada data numerik YouTube Studio tanpa melibatkan aspek temporal (waktu publikasi) atau tekstual (judul, komentar). Dengan demikian, generalisasi model ke channel lain perlu diuji lebih lanjut sebelum implementasi skala luas.

Validasi Silang (5-Fold)

Berdasarkan hasil silang (5-fold) maka didapatkan rata-rata CV Score 98,40%; Standar Deviasi 0,0196 (1,96%); dan Min – Max 96,00% – 100,00%. Nilai validasi silang yang konsisten di atas 96% dengan standar deviasi yang rendah mengonfirmasi bahwa model Random Forest yang dibangun memiliki stabilitas dan keandalan yang tinggi. Artinya, performa model tidak bergantung secara berlebihan pada satu partisi data tertentu.

Analisis Fitur Penting

Untuk memahami faktor apa saja yang paling berpengaruh terhadap keputusan klasifikasi, dilakukan analisis *feature importance* menggunakan *Gini importance*, sebagaimana ditampilkan pada Tabel 10.

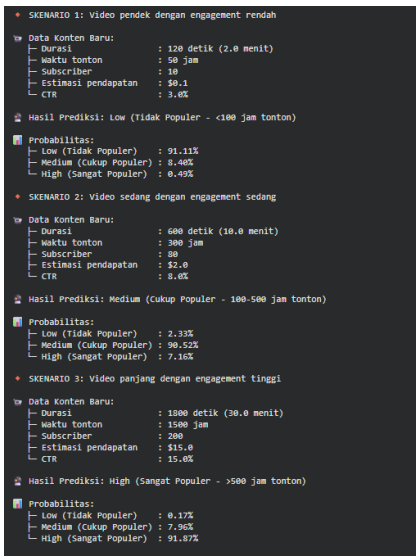
Tabel 10. *Feature Importance* (*Gini Importance*)

Peringkat	Nama Fitur	Nilai Importance	Persentase
1	Waktu tonton (jam)	0,4637	46,37%
2	Subscriber	0,2089	20,89%
3	Rasio klik-tayang (%)	0,1884	18,84%
4	Durasi	0,0916	9,16%
5	Estimasi pendapatan (USD)	0,0473	4,73%

Berdasarkan Tabel 8, fitur Waktu tonton (jam) menjadi prediktor paling dominan dengan kontribusi hampir separuh dari keseluruhan kepentingan fitur (46,37%), diikuti oleh Subscriber (20,89%) dan Rasio klik-tayang (18,84%). Temuan ini mengindikasikan bahwa durasi menonton dan basis pelanggan merupakan faktor kunci dalam menentukan tingkat popularitas suatu video, sementara durasi video dan estimasi pendapatan memiliki pengaruh yang relatif lebih kecil.

Pengujian pada Skenario Konten Baru

Untuk menguji kemampuan prediksi model dalam kondisi dunia nyata yang beragam, model diuji pada tiga skenario konten baru yang mewakili tingkat popularitas rendah, sedang, dan tinggi. Hasil prediksi skenario konten baru disajikan pada Gambar 4.



Gambar 4. Hasil Prediksi

Gambar 4 menyajikan hasil uji coba model Random Forest terhadap tiga skenario konten video baru yang mewakili tingkat popularitas berbeda: rendah (Skenario 1), sedang (Skenario 2), dan tinggi (Skenario 3). Model mampu memprediksi ketiga skenario dengan tingkat keyakinan (*probabilitas*) di atas 90% pada masing-masing kelas. Skenario 1 diprediksi sebagai kelas Low dengan probabilitas 91,11%, Skenario 2 sebagai kelas *Medium* dengan probabilitas 90,52%, dan Skenario 3 sebagai kelas *High* dengan probabilitas 91,87%. Hasil ini menunjukkan bahwa model Random Forest telah belajar pola hubungan antara fitur (durasi, waktu tonton, subscriber, pendapatan, CTR) dengan tingkat popularitas secara general, tidak sekadar menghafal data latih, sehingga layak digunakan untuk memprediksi konten baru di masa

mendatang. Namun, perlu dicatat bahwa skenario ini bersifat simulasi; validasi lapangan pada konten yang benar-benar dipublikasikan masih diperlukan untuk memastikan keandalan model di dunia nyata.

Ringkasan Performa Model

Berdasarkan seluruh hasil evaluasi yang telah dipaparkan, ringkasan performa model Random Forest disajikan pada Tabel 11.

Tabel 11. Ringkasan Final Performa Model Random Forest

Komponen	Metrik	Nilai
Split Data	Training : Testing	75 : 50 (60% : 40%)
Akurasi	Training Accuracy	100,00%
Testing Accuracy	98,00%	
Gap	2,00%	
Metrik Evaluasi (Weighted)	Precision	96,97%
Recall	91,67%	
F1-Score	97,90%	
Cross Validation (5-Fold)	Mean CV Score	98,40%
Std Deviasi	1,96%	
Feature Importance (Top-3)	1. Waktu tonton	46,37%
	2. Subscriber	**20,89%**
	3. Rasio klik-tayang	**18,84%**

Secara keseluruhan, model Random Forest yang dibangun menunjukkan performa sangat baik (akurasi uji 98%, *F1-score* tertimbang 97,9%) dan stabil (validasi silang 98,4% ± 1,96%) dalam memprediksi popularitas video pada dataset Blora TV. Hasil validasi silang 5-lipat menghasilkan rata-rata skor 98,40%, menegaskan stabilitas model. Berdasarkan *feature importance*, tiga fitur paling berpengaruh adalah Waktu tonton (46,37%), Subscriber (20,89%), dan Rasio klik-tayang (18,84%). Temuan ini juga mengonfirmasi bahwa waktu tonton dan jumlah subscriber merupakan metrik paling kunci dalam menentukan keberhasilan sebuah konten video.

KESIMPULAN

Berdasarkan hasil penelitian, model Random Forest terbukti sangat efektif dan stabil dalam mengklasifikasikan *engagement rate* konten video Blora TV dengan akurasi *testing* 98% dan F1-score 97,9%, serta gap akurasi hanya 2% yang mengindikasikan tidak terjadi *overfitting* signifikan. Waktu tonton (46,37%), subscriber (20,89%), dan CTR (18,84%) merupakan faktor paling dominan dalam menentukan popularitas video. Meskipun model unggul pada kelas *Low* dan *Medium*, kelas *High* masih memiliki recall 75% akibat ketidakseimbangan kelas (hanya 7,2% video populer). Keterbatasan utama penelitian ini meliputi: (1) ukuran dataset yang kecil (125 video) sehingga generalisasi terbatas, (2) dominasi kelas *Low* (71,2%) yang menyulitkan deteksi video populer, dan (3) fitur yang hanya terbatas pada data numerik tanpa aspek temporal atau tekstual.

Sebagai saran, penelitian selanjutnya perlu memperluas dataset, menambahkan fitur waktu publikasi dan analisis thumbnail, serta melakukan validasi lapangan sebelum implementasi skala luas. Serta menambahkan fitur tekstual dari judul video dan analisis sentimen komentar, serta mengembangkan model *time series* untuk menangkap pola musiman. Bagi Blora TV, hasil ini dapat dijadikan acuan strategis untuk memprioritaskan konten yang

mampu mempertahankan waktu tonton penonton dan mendorong pertumbuhan subscriber melalui call-to-action yang efektif.

DAFTAR PUSTAKA

[1] K. L. Tan, C. P. Lee, and K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Appl. Sci.*, vol. 13, no. 7, 2023, doi: 10.3390/app13074550.

[2] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decis. Anal. J.*, vol. 3, no. May, p. 100073, 2022, doi: 10.1016/j.dajour.2022.100073.

[3] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P. M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst. Appl.*, vol. 223, no. August 2022, 2023, doi: 10.1016/j.eswa.2023.119862.

[4] M. A. M. Setiawan, K. Kusrini, and A. D. Hartono, "Menggunakan Metode Machine Learning Untuk Memprediksi Nilai Mahasiswa Dengan Model Prediksi Multiclass," *J. Inform. J. Pengemb. IT*, vol. 10, no. 1, pp. 190–204, 2025, doi: 10.30591/jpit.v10i1.8334.

[5] R. Alfian, M. Shobari, H. Santoso, E. D. Ulya, and R. T. Harpin, "Strategi Konten Edutainment Instagram Terhadap Ketertarikan Pembelian Anggota DWDG IPB," Jurnal Riset Komunikasi dan Media, vol. 5, no. 2, pp. 4177–4191, 2026.

[6] S. H. Fikriyah, P. B. Sulistyo, and I. Tomohardjo, "Strategi Peningkatan Engagement Melalui Produksi Konten dan Preferensi Audiens," *J. Din. Ilmu Komun.*, vol. 11, no. 2, pp. 215–229, 2025, doi: 10.32509/dinamika.v11i2.5465.

[7] N. I. T. Ati, F. A. Fadilah, and T. W. Nurdiani, "Studi Kasus Customer Engagement dan Content Planning dalam Manajemen Pemasaran PT Akademi Smart Indonesia," Jurnal Indonesian Management, vol. 5, no. 3, pp. 1–15, 2025.

[8] F. Widianto, "Analisis Sentimen Komentar Youtube tentang Konflik Iran-Israel Menggunakan Orange Data Mining," *Sains Data J. Stud. Mat. dan Teknol.*, vol. 3, no. 2, pp. 81–88, 2025, doi: 10.52620/sainsdata.v3i2.278.

[9] M. A. Al Montaser *et al.*, "Sentiment analysis of social media data: Business insights and consumer behavior trends in the USA," *Edelweiss Appl. Sci. Technol.*, vol. 9, no. 1, pp. 545–565, 2025, doi: 10.55214/25768484.v9i1.4164.

[10] M. Fauzan and R. Ahmad, "Program Kampus Merdeka Berbasis Web Sentiment Analysis of Youtube Comments About Program Using Naïve Bayes Multinomial Algorithm," in *Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, Jakarta, Indonesia, Sep. 2023, pp. 864–871. [Online]. Available: https://senafti.budiluhur.ac.id/index.php/senafti/article/view/929

[11] M. Hussein Umar, R. Daud Antony Pangaribuan, R. Primadana, I. Budiawan, and R. Pakpahan, "Penerapan Algoritma Random Forest Untuk Prediksi Tingkat Stres Dari Aktivitas Media Sosial," Jurnal Media Informatika (JUMIN), vol. 6, no. 6, pp. 3113–3122, 2025.

[12] F. S. S. Nagalay et al., "Penerapan Algoritma Random Forest dengan Pendekatan Hybrid Feature Engineering untuk Klasifikasi Spam Judi Online," Jurnal Ilmiah Teknologi Informasi, vol. 8, no. 1, pp. 45–55, 2024.

[13] A. M. Wahid, T. Turino, K. A. Nugroho, T. S. Maharani, D. Darmono, and F. S. Utomo, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis

TF-IDF," *Jurnal Pendidikan dan Teknologi Indonesia*, vol. 4, no. 8, pp. 381–392, 2024. [Online]. Available: https://jpti.journals.id/index.php/jpti/article/view/602

[14] S. Wei, K. Shores, and Y. Xu, "A Comparison of Machine Learning-Based Approaches in Estimating Surface PM2.5 Concentrations Focusing on Artificial Neural Networks and High Pollution Events," *Atmosphere (Basel).*, vol. 16, no. 1, 2025, doi: 10.3390/atmos16010048.

[15] E. Ismanto, A. G. Dalimunthe, M. Iqbal, and F. Addinunnisa, "Analisis Komparatif Model Machine Learning dan Deep Learning pada Peramalan Harga Saham Time Series," vol. 2, no. 1, pp. 131–139, 2026.

[16] H. Eldo, A. Ayuliana, D. Suryadi, G. Chrisnawati, and L. Judijanto, "Penggunaan Algoritma Support Vector Machine (SVM) Untuk Deteksi Penipuan pada Transaksi Online," *J. Minfo Polgan*, vol. 13, no. 2, pp. 1627–1632, 2024, doi: 10.33395/jmp.v13i2.14186.

[17] R. A. Silvana, N. Anggiani, A. Labib, and R. L. Pratiwi, "Perbandingan Kinerja Algoritma Decision Tree dan Random Forest dalam Memprediksi Kepuasan Penumpang Maskapai," Jurnal Ilmiah Teknologi dan Informasi, vol. 5, no. 1, pp. 55–65, Sep. 2025.

[18] F. Cappelli, G. Castronuovo, S. Grimaldi, and V. Telesca, "Random Forest and Feature Importance Measures for Discriminating the Most Influential Environmental Factors in Predicting Cardiovascular and Respiratory Diseases," *Int. J. Environ. Res. Public Health*, vol. 21, no. 7, 2024, doi: 10.3390/ijerph21070867.

[19] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.

[20] O. O. Bifarin, "Interpretable machine learning with treebased shapley additive explanations: Application to metabolomics datasets for binary classification," *PLoS One*, vol. 18, no. 5 May, 2023, doi: 10.1371/journal.pone.0284315.