

Klasifikasi Performa Video Free Fire di YouTube Menggunakan Pelebelan Otomatis Berbasis K-Means, Random Forest, dan SHAP

Wayan Praka ^{*1} dan Erik Iman Heri Ujjianto ²

¹ Universitas Teknologi Yogyakarta; wayanpraka123@gmail.com

² Universitas Teknologi Yogyakarta; erik.iman@uty.ac.id

Abstrak: Perkembangan platform YouTube mendorong kreator konten Free Fire untuk memahami faktor-faktor yang memengaruhi performa video untuk meningkatkan strategi publikasi yang lebih efektif. Namun, penelitian yang mengintegrasikan pelebelan otomatis, klasifikasi, dan interpretabilitas model untuk mengungkap faktor penentu masih terbatas. Penelitian ini bertujuan mengembangkan model klasifikasi performa video menggunakan pelebelan otomatis berbasis K-Means, algoritma Random Forest, dan Explainable Artificial Intelligence berbasis SHAP. Dataset yang digunakan terdiri atas 4.000 video Free Fire dari kanal YouTube Indonesia. Proses pelabelan menggunakan algoritma K-Means berdasarkan evaluasi Elbow Method, Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index. Hasil evaluasi menunjukkan bahwa K=2 memberikan kualitas kluster terbaik dengan nilai Silhouette Score tertinggi, dipilih untuk membentuk dua label, yaitu High Performance sebesar 74,23% dan Low Performance sebesar 25,77%. Selanjutnya, model Random Forest yang dioptimalkan dengan GridSearchCV serta pembagian 80% data pelatihan dan 20% data pengujian. Hasil dari pengujian Random Forest mencapai akurasi sebesar 92,37% dan akurasi 10-Fold Cross Validation rata-rata sebesar 91,59%. Analisis SHAP menunjukkan bahwa Subscriber_Channel merupakan faktor yang paling berpengaruh, diikuti oleh Duration_Sec, Title_Length, Video_Age_Days, Upload_Hour_WIB dan Upload_day_WIB. Kebaruan penelitian ini terletak pada integrasi pelebelan otomatis berbasis K-Means, optimasi Random Forest menggunakan GridSearchCV, serta interpretasi model menggunakan SHAP dalam analisis performa video YouTube. Penelitian ini masih terbatas pada video Free Fire dari kanal YouTube Indonesia sehingga generalisasi terhadap kategori konten maupun wilayah lain memerlukan penelitian lanjutan.

Keywords: Popularitas Video YouTube, Free Fire, K-Means Clustering, Random Forest, SHapley Additive exPlanations.

DOI: <https://doi.org/10.47134/jacis.v6i2.203>

*Correspondensi: Wayan Praka

Email: wayanpraka123@gmail.com

Receive: 24 Juni 2026

Accepted: 10 Juli 2026

Published: 17 Juli 2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Abstract: The rapid growth of YouTube has encouraged Free Fire content creators to better understand the factors influencing video performance to develop more effective publishing strategies. However, studies integrating automatic labeling, classification, and model interpretability to identify the key determinants of video performance remain limited. This study proposes a video performance classification framework by integrating K-Means-based automatic labeling, the Random Forest algorithm, and Explainable Artificial Intelligence (XAI) using SHAP. The dataset consisted of 4,000 Free Fire videos collected from Indonesian YouTube channels. Automatic labeling was performed using the K-Means clustering algorithm, with the optimal number of clusters determined through the Elbow Method, Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index. The

evaluation results indicated that K=2 produced the best clustering quality, achieving the highest Silhouette Score and generating two performance labels: High Performance (74.23%) and Low Performance (25.77%). Subsequently, a Random Forest classifier was optimized using GridSearchCV and evaluated with an 80:20 train-test split. The proposed model achieved an accuracy of 92.37% on the testing dataset and an average accuracy of 91.59% under 10-fold cross-validation. SHAP analysis revealed that Subscriber_Channel was the most influential feature, followed by Duration_Sec, Title_Length, Video_Age_Days, Upload_Hour_WIB, and Upload_Day_WIB. The novelty of this study lies in the integration of K-Means based automatic labeling, GridSearchCV optimized Random Forest classification, and SHAP based model interpretation for YouTube video performance analysis. Nevertheless, this study is limited to Free Fire videos from Indonesian YouTube channels; therefore, further research is required to evaluate the proposed framework across different content categories and geographical regions.

Keywords: YouTube Video Popularity, Free Fire, K-Means Clustering, Random Forest, SHapley Additive exPlanations.

PENDAHULUAN

Perkembangan platform berbagi video, khususnya YouTube, telah mendorong pertumbuhan budaya digital sekaligus menghasilkan *digital trace data* yang bermanfaat untuk menganalisis perilaku pengguna dan penyebaran konten di media sosial [1], [2], [3]. Di Indonesia, konten *gaming*, terutama Free Fire, menjadi salah satu kategori dengan tingkat konsumsi dan keterlibatan audiens yang tinggi sehingga mendorong kreator menghasilkan berbagai konten, seperti *gameplay*, strategi permainan, dan hiburan [4]. Di tengah persaingan algoritma rekomendasi YouTube, pemahaman terhadap faktor-faktor yang memengaruhi performa video menjadi penting untuk menyusun strategi publikasi yang lebih efektif [5]. Performa video umumnya diukur melalui metrik seperti *views*, *likes*, *comments*, dan *engagement*, namun dipengaruhi oleh berbagai faktor yang kompleks sehingga pendekatan *Machine Learning* semakin banyak dimanfaatkan untuk memodelkan pola penyebaran konten dan memprediksi respons audiens [6], [7], [8], [9].

Salah satu tantangan dalam analisis performa konten media sosial adalah tidak tersedianya label performa secara langsung. Kondisi ini mendorong pemanfaatan metode *unsupervised learning*, seperti K-Means Clustering, untuk membentuk kelompok data berdasarkan karakteristik yang serupa [10], [11], [12], [13]. Hasil pengelompokan selanjutnya dapat dimanfaatkan pada proses klasifikasi menggunakan Random Forest, yang dikenal memiliki kemampuan tinggi dalam menangani data tabular dan mengurangi *overfitting* [14], [15], [16]. Selain menghasilkan prediksi yang akurat, perkembangan *Explainable Artificial Intelligence* (XAI), khususnya SHAP, memungkinkan interpretasi kontribusi setiap fitur terhadap keputusan model sehingga meningkatkan transparansi hasil klasifikasi [17], [18].

Meskipun berbagai penelitian telah menunjukkan performa yang baik, masih terdapat beberapa keterbatasan. Sebagian besar penelitian menggunakan pelabelan manual yang berpotensi menimbulkan subjektivitas, belum memvalidasi kualitas label yang dihasilkan, serta hanya mengevaluasi satu algoritma klasifikasi tanpa membandingkan performanya dengan metode lain. Selain itu, penelitian terdahulu lebih berfokus pada peningkatan akurasi daripada interpretasi faktor-faktor yang memengaruhi hasil prediksi, padahal informasi tersebut penting bagi pengambilan keputusan berbasis data. Di sisi lain, Random Forest masih bersifat *black box*, sehingga memerlukan pendekatan interpretabilitas seperti SHAP untuk meningkatkan transparansi model [17], [18], [19].

Berdasarkan kesenjangan tersebut, penelitian ini mengusulkan suatu klasifikasi performa video *Free Fire* di *YouTube* yang mengintegrasikan pelabelan otomatis berbasis *K-Means*, validasi kualitas kluster menggunakan *Elbow Method*, *Silhouette Score*, *Calinski-Harabasz Index*, *Davies-Bouldin Index*, perbandingan enam algoritma klasifikasi, yaitu *Decision Tree*, *Logistic Regression*, *Support Vector Machine*, *XGBoost*, *LightGBM*, dan *Random Forest*. *Random Forest* menggunakan *GridSearchCV*, serta interpretasi model menggunakan SHAP. Penelitian ini tidak hanya menghasilkan model klasifikasi yang memiliki performa prediksi yang baik, tetapi juga menyediakan penjelasan yang transparan mengenai faktor-faktor yang memengaruhi performa video *Free Fire* di *YouTube*. Temuan penelitian ini diharapkan dapat menjadi dasar pengambilan keputusan berbasis data bagi kreator konten dalam menyusun strategi publikasi video yang lebih efektif serta memberikan kontribusi terhadap pengembangan penelitian analitik media sosial berbasis *Explainable Artificial Intelligence*.

METODE

Penelitian ini membangun model klasifikasi performa video *Free Fire* di *YouTube* menggunakan pendekatan *K-Means Clustering*, *Random Forest*, dan *Explainable Artificial Intelligence (XAI)* berbasis SHAP. Tahapan penelitian meliputi pengumpulan data, *preprocessing*, transformasi dan normalisasi data, pembentukan label menggunakan *K-Means*, klasifikasi menggunakan *Random Forest*, evaluasi model, serta interpretasi hasil menggunakan SHAP.

Pengumpulan Data

Data dikumpulkan menggunakan *YouTube Data API v3* melalui *Python* pada *Google Colab* selama periode 13 Juni 2025–14 Juni 2026. Pengambilan data dilakukan menggunakan berbagai kata kunci yang merepresentasikan konten *Free Fire* di Indonesia dengan parameter *regionCode=ID*, *relevanceLanguage=id*, dan rentang waktu publikasi sesuai periode penelitian. Selanjutnya dilakukan penyaringan berdasarkan kriteria video yang membahas *Free Fire*, berasal dari kanal *YouTube Indonesia*, memiliki metadata lengkap, dipublikasikan pada periode pengamatan, serta berdurasi minimal 180 detik sehingga video *YouTube Shorts* tidak disertakan. Data duplikat dihapus berdasarkan atribut *Video_ID*.

Dataset akhir terdiri atas 4.000 video dari 464 kanal *YouTube Indonesia* yang direpresentasikan menggunakan 12 variabel, yaitu *Video_ID*, *Title*, *Published_Date_WIB*, *Upload_Hour_WIB*, *Upload_Day_WIB*, *Video_Age_Days*, *Duration_Sec*, *Subscriber_Channel*, *View_Count*, *Like_Count*, *Comment_Count*, dan *Engagement_Rate* [1], Sampel dataset disajikan pada Tabel 1.

Tabel 1. Sampel Dataset Video

Nama variabel	Video 1	Video 2	Video 3
Video_ID	rM8ux30Imos	d8lZWTa-o9M	bqTo78EnoYk
Title	Event Baru Pildun Epep	Akhimya Gw Main Epep Lagi	Toko Kelontong- Free Fire
Published_Date_WIB	03/06/2026 09:00	21/11/2025 19:17	20/02/2026 03:00
Upload_Hour_WIB	9	19	3
Upload_Day_WIB	Wednesday	Friday	Friday
Video_Age_Days	13	207	116
Duration_Sec	482	1.544	484
Subscriber_Channel	937.000	12.100.000	937.000
View_Count	54.574	327.924	140.038
Like_Count	1.739	9.129	4.259
Comment_Count	462	802	2.077
Engagement_Rate	40.330	30.280	45.240

Preprocessing Data

Tahap *preprocessing* dilakukan untuk meningkatkan kualitas data sebelum proses pemodelan. Tahapan ini meliputi penghapusan data duplikat, penanganan *missing value*, deteksi *outlier* menggunakan metode Interquartile Range (IQR), serta validasi tipe data. Selanjutnya dilakukan transformasi logaritmik pada variabel View_Count, Like_Count, dan Comment_Count untuk mengurangi kemencengan distribusi data [5]. Kemudian seluruh fitur numerik dinormalisasi menggunakan StandardScaler agar memiliki skala yang sama sebelum proses clustering [11], [20].

K-Means Clustering

K-Means Clustering digunakan untuk membentuk label performa video berdasarkan karakteristik interaksi audiens [10] [11]. Jumlah klaster optimum ditentukan menggunakan kombinasi Elbow Method, Silhouette Score, Calinski-Harabasz Index, dan Davies-Bouldin Index [20], [13]. Hasil evaluasi keempat indeks tersebut digunakan untuk memilih jumlah klaster yang memberikan kualitas pengelompokan terbaik.

Klasifikasi Random Forest

Label hasil K-Means selanjutnya digunakan sebagai target klasifikasi menggunakan Random Forest [14]. Algoritma ini dipilih karena mampu menangani hubungan nonlinier antarfitur serta memiliki ketahanan terhadap *overfitting* [16]. *Random Forest* memproses dengan cara membentuk sekumpulan pohon keputusan secara acak selama fase pelatihan, lalu menentukan kelas prediksi akhir berdasarkan mekanisme pemungutan suara terbanyak [21]. Fitur yang digunakan sebagai prediktor meliputi Title_Length, Video_Age, Duration, Subscriber_Channel, dan Upload_Hour.

Evaluasi Model

Performa model dievaluasi menggunakan Confusion Matrix [16], [22] dengan metrik **Accuracy**, **Precision**, **Recall**, dan **F1-score** untuk mengukur kemampuan model dalam mengklasifikasikan video ke dalam kategori performa yang sesuai.

SHAP Explainability

Interpretasi model dilakukan menggunakan SHapley Additive exPlanations (SHAP) [18], [17] untuk mengukur kontribusi masing-masing fitur terhadap hasil prediksi Random Forest. Analisis dilakukan pada fitur Title_Length, Upload_Hour_WIB, Upload_Day_WIB,

Video_Age_Days, Duration_Sec, dan Subscriber_Channel sehingga proses pengambilan keputusan model dapat dijelaskan secara lebih transparan.

HASIL DAN PEMBAHASAN

Deskripsi Dataset

Setelah melalui tahap *preprocessing*, diperoleh 4.000 baris data video *Free Fire* yang layak dianalisis. Analisis statistik deskriptif dilakukan terhadap sepuluh variabel yang merepresentasikan karakteristik konten, atribut saluran, dan keterlibatan *audiens* untuk memberikan gambaran mengenai pola serta distribusi data. Ringkasan hasil analisis disajikan pada tabel 2.

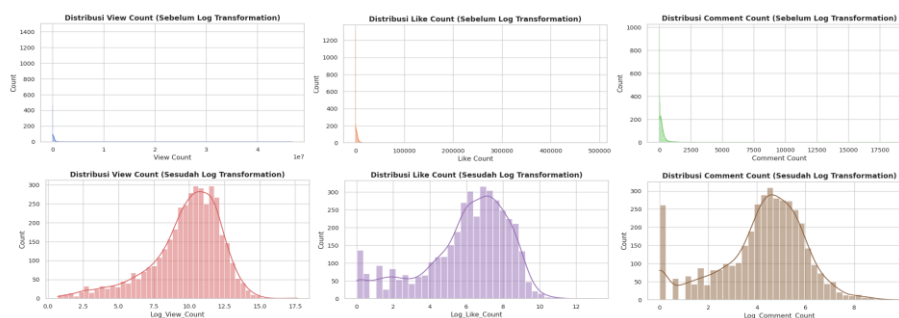
Tabel 2. Deskriptif Variabel Dataset

Variabel	Min	Max	Mean	Std Deviasi
Title Length	11	110,0	82,04	18,47
Upload Hour WIB	0	23,0	15,12	4,27
Upload Day WIB	0	6,0	3,05	1,98
Video Age Day	1	369,0	142,41	101,48
Duration Sec	181	2.770,0	900,41	436,21
Subscriber Channel	0	113.000.000,0	1.862.798,00	3.883.200,00
View Count	1	46.790.490,0	124.679,80	946.350,80
Like Count	0	491.046,0	2.005,82	8.516,27
Comment Count	0	19.022,0	222,19	596,68
Engagement Rate	0	724.140,0	20.328,52	31.046,26

Berdasarkan Tabel 2 diketahui bahwa dataset memiliki karakteristik yang cukup beragam, yang ditunjukkan oleh rentang nilai minimum dan maksimum yang lebar pada hampir seluruh variabel. Variabel *Video_Age_Days* memiliki rentang 1-369 hari dengan rata-rata 142,41 hari, menunjukkan bahwa dataset mencakup video yang baru dipublikasikan hingga video satu tahun. Rata-rata *Duration_Sec* dari 900,41 detik mengindikasikan bahwa mayoritas video merupakan video berdurasi panjang dan tidak termasuk kategori *YouTube Shorts*. Variabel *Subscriber_Channel* menunjukkan variasi yang sangat tinggi, dengan nilai minimum 0 dan maksimum 113.000.000 pelanggan. Kondisi serupa juga terlihat pada variabel *View_Count*, *Like_Count*, *Comment_Count*, dan *Engagement_Rate*, yang memiliki rentang nilai sangat lebar disertai standar deviasi tinggi. Hal tersebut mengindikasikan bahwa distribusi data cenderung tidak merata, di mana sebagian kecil video memperoleh performa yang sangat tinggi, sedangkan sebagian besar video memiliki tingkat performa yang relatif lebih rendah. Karakteristik ini merupakan pola yang umum dijumpai pada data media sosial dan platform berbagi video, sehingga diperlukan tahapan transformasi logaritmik serta standardisasi sebelum proses *clustering* untuk mengurangi pengaruh perbedaan skala antarvariabel dan meningkatkan stabilitas pembentukan cluster.

Hasil Transformasi Logaritmik

Transformasi *logaritmik* diterapkan pada atribut dengan rentang nilai yang tinggi untuk menstabilkan *varians*, mengurangi pengaruh nilai ekstrem, dan memperbaiki distribusi data agar lebih mendekati normal. Transformasi Logaritmik divisualisasikan pada gambar 1.

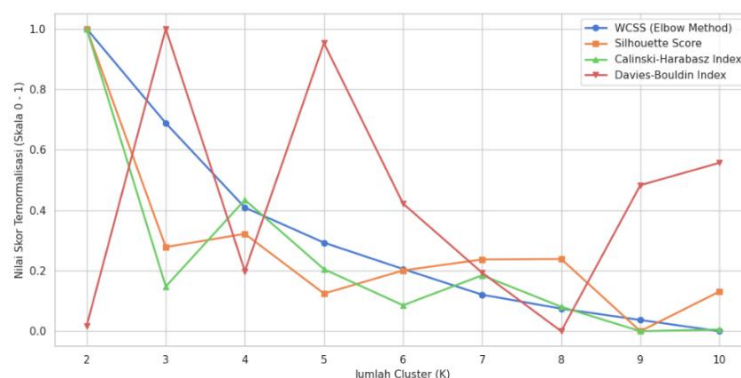


Gambar 1. Distribusi Data Sebelum dan Sesudah Transformasi Logaritmik

Pada Gambar 1, Data sebelum transformasi logaritmik menunjukkan distribusi yang sangat menceng ke kanan atau *positively skewed*, dengan konsentrasi nilai pada rentang rendah dan keberadaan nilai ekstrem pada rentang tinggi. Setelah transformasi *logaritmik* diterapkan, distribusi data menjadi lebih seimbang dan mendekati normal karena tingkat *skewness* berkurang serta sebaran data menjadi lebih merata. Perbaikan distribusi ini berdampak positif pada proses *clustering* dengan mengurangi dominasi nilai ekstrem dalam perhitungan jarak, sehingga *centroid* yang terbentuk menjadi lebih representatif serta menghasilkan proses *clustering* yang lebih stabil dan valid untuk analisis lanjutan.

Hasil K-Means Clustering

Algoritma *K-Means* menetapkan jumlah kluster (k) sebelum proses pengelompokan data dijalankan. Pemilihan nilai k dievaluasi menggunakan empat pendekatan yaitu *Within Cluster Sum of Squares* (WCSS) melalui *Elbow Method*, *Silhouette Score*, *Calinski-Harabasz Index*, dan *Davies-Bouldin Index*. Sebagai ukuran kualitas pemisahan antar kluster, yang dapat divisualisasikan pada gambar 1.



Gambar 2. Grafik *Elbow Method*, *Silhouette Score*, *Calinski-Harabasz Index* dan *Davies-Bouldin Index*

Pada Gambar 2, Penentuan jumlah cluster dilakukan dengan mengevaluasi nilai K pada rentang 2 hingga 10. kurva WCSS mengalami penurunan yang sangat tajam dari $K=2$ menuju $K=3$, kemudian mulai melandai pada nilai K berikutnya. Pola tersebut menunjukkan bahwa penambahan jumlah cluster setelah $K=2$ hanya memberikan penurunan variasi dalam cluster yang relatif kecil. Hasil tersebut diperkuat oleh *Silhouette Score*, yang mencapai nilai tertinggi pada $K=2$ sebesar 0,524, menunjukkan bahwa objek dalam setiap cluster memiliki tingkat kemiripan yang tinggi terhadap clusternya sendiri dan memiliki pemisahan yang baik terhadap cluster lainnya. Selain itu, *Calinski-Harabasz Index* juga memperoleh nilai maksimum pada $K=2$, sedangkan *Davies-Bouldin Index* mencapai nilai minimum pada jumlah cluster yang sama. Kombinasi keempat indikator tersebut menunjukkan bahwa $K=2$ merupakan jumlah

cluster yang memberikan kualitas pemisahan terbaik dibandingkan alternatif jumlah cluster lainnya. Oleh karena itu, penelitian ini menetapkan dua cluster sebagai dasar pembentukan label otomatis.

Tabel 3. Karakteristik Cluster Hasil *K-Means*

Variabel	Cluster 0	Cluster 1
Jumlah Video	1.031	2.969
Mean View_Count	2.874,72	166.977,16
Mean Like_Count	41,16	2.688,06
Mean Comment_Count	11,81	295,25
Mean Engagement_Rate	29.423,84	17.170,13

Pada Tabel 3, setelah jumlah cluster optimum ditentukan, algoritma *K-Means* diterapkan menggunakan empat variabel, yaitu *Log_View_Count*, *Log_Like_Count*, *Log_Comment_Count*, dan *Engagement_Rate*. Transformasi logaritmik pada tiga variabel bertujuan mengurangi kemencengan distribusi sehingga proses clustering menjadi lebih stabil. Hasil dari clustering menjadi dua kelompok, yaitu Cluster 1 sebanyak 2.969 video dan Cluster 0 sebanyak 1.031 video. Berdasarkan nilai rata-rata setiap variabel, Cluster 1 memiliki *View_Count*, *Like_Count*, dan *Comment_Count* yang lebih tinggi dibandingkan Cluster 0, sedangkan *Engagement_Rate* pada Cluster 0 lebih tinggi. Hal ini menunjukkan bahwa video dengan jumlah penonton lebih sedikit cenderung memiliki proporsi interaksi yang lebih tinggi, sedangkan video dengan jangkauan *audience* lebih luas menghasilkan volume interaksi yang lebih besar meskipun tingkat *engagement* relatif lebih rendah. Temuan ini mengindikasikan bahwa *K-Means* membentuk cluster berdasarkan kombinasi beberapa indikator performa, sehingga menghasilkan pengelompokan yang lebih representatif terhadap karakteristik video.

Tabel 4. Hasil Validasi Cluster

Parameter	Hasil
Silhouette Score	0,524
Shapiro-Wilk (p-value)	$1,5114 \times 10^{-88}$
Mann-Whitney U Test (p-value)	<0,001

Pada tabel 4, memastikan bahwa cluster yang terbentuk benar-benar merepresentasikan dua kelompok video yang berbeda, dilakukan proses validasi menggunakan evaluasi internal dan pengujian statistik. Nilai *Silhouette Score* sebesar 0,524 menunjukkan bahwa kualitas pemisahan antarcluster berada pada kategori baik. Selanjutnya dilakukan uji normalitas *Shapiro-Wilk* terhadap distribusi data. Hasil pengujian menghasilkan *p-value* sebesar $1,5114 \times 10^{-88}$, sehingga data tidak memenuhi asumsi distribusi normal. Oleh karena itu, pengujian perbedaan antarcluster dilanjutkan menggunakan *Mann-Whitney U Test* sebagai uji nonparametrik. Hasil *Mann-Whitney U Test* menghasilkan *p-value* < 0,001, yang menunjukkan adanya perbedaan yang signifikan antara kedua cluster. Dengan demikian, cluster yang terbentuk tidak terjadi secara acak, tetapi merepresentasikan dua kelompok video dengan karakteristik performa yang berbeda secara statistik.

Tabel 5. Distribusi Label Hasil Pelebelan Otomatis

Label	Jumlah	Persentase
High Performance	2.969	74,23%
Low Performance	1.031	25,77%
Total	4.000	100%

Pada Tabel 5, hasil clustering dan validasi selanjutnya diinterpretasikan menjadi dua label, yaitu *High Performance* dan *Low Performance*, berdasarkan karakteristik gabungan empat indikator performa yaitu *View_Count*, *Like_Count*, *Comment_Count*, dan *Engagement_Rate*.

Cluster dengan rata-rata *View_Count*, *Like_Count*, dan *Comment_Count* yang lebih tinggi ditetapkan sebagai *High Performance*, sedangkan cluster lainnya sebagai *Low Performance*. Hasil pelabelan menunjukkan bahwa 2.969 video atau 74,23% termasuk kategori *High Performance*, sedangkan 1.031 video atau 25,77% termasuk *Low Performance*. Label tersebut bukan merupakan label aktual dari *YouTube* maupun hasil pelabelan manual, melainkan *pseudo-label* yang dibentuk secara otomatis oleh algoritma *K-Means* berdasarkan karakteristik multivariat data. Validitasnya didukung oleh hasil evaluasi menggunakan empat indeks clustering serta pengujian statistik yang menunjukkan perbedaan signifikan antarkluster. Dengan demikian, label yang dihasilkan dapat digunakan sebagai target klasifikasi untuk merepresentasikan performa relatif video *Free Fire* dalam dataset penelitian ini.

Hasil Klasifikasi Random Forest

Sebelum menentukan model klasifikasi utama, dilakukan evaluasi terhadap enam algoritma machine learning, yaitu *Decision Tree*, *Logistic Regression*, *Support Vector Machine*, *XGBoost*, *LightGBM*, dan *Random Forest*. Seluruh algoritma dilatih menggunakan dataset yang sama dengan pembagian 80% data pelatihan dan 20% data pengujian serta dievaluasi menggunakan metrik *Accuracy*, *Precision*, *Recall*, dan *F1-score*. Tahapan ini bertujuan memperoleh model dengan performa terbaik sekaligus memastikan bahwa pemilihan *Random Forest* didasarkan pada hasil perbandingan yang objektif.

Tabel 6. Perbandingan Performa Model Klasifikasi

Algoritma	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Decision Tree	88,63	92,70	91,92	92,31
Logistic Regression	81,88	83,36	94,44	88,56
Support Vector Machine	79,88	81,15	94,95	87,51
XGBoost	91,88	94,16	94,95	94,55
LightGBM	92,00	93,87	95,45	94,66
Random Forest	92,50	95,10	94,78	94,94

Pada Tabel 6, algoritma berbasis *ensemble learning* yaitu *Random Forest*, *LightGBM*, dan *XGBoost*, menunjukkan performa yang lebih baik dibandingkan algoritma klasifikasi konvensional. *Random Forest* memperoleh *Accuracy* tertinggi sebesar 92,50%, diikuti oleh *LightGBM* sebesar 92,00% dan *XGBoost* sebesar 91,88%. Selain memiliki akurasi tertinggi, *Random Forest* juga menghasilkan nilai *Precision* sebesar 95,10% dan *F1-score* sebesar 94,94%, yang menunjukkan kemampuan model dalam mengklasifikasikan video *High Performance* dan *Low Performance* secara lebih seimbang. Sebaliknya, algoritma *Decision Tree* memperoleh akurasi sebesar 88,63%, sedangkan *Logistic Regression* dan *Support Vector Machine* masing-masing hanya mencapai 81,88% dan 79,88%. Hasil tersebut menunjukkan bahwa model linier maupun pohon keputusan tunggal kurang mampu menangkap hubungan nonlinier antarvariabel metadata video dibandingkan algoritma *ensemble*. Oleh karena itu, *Random Forest* dipilih sebagai model utama untuk tahap optimasi *hyperparameter* dan analisis interpretabilitas menggunakan *SHAP*.

Tabel 7. Hyperparameter Random Forest

Parameter	criterion	max_depth	max_features	min_samples_leaf	min_samples_split	n_estimators
Nilai Terbaik	entropy	20	sqrt	1	2	100

Pada tabel 7, *hyperparameter* menggunakan *GridSearchCV* dengan skema *5-Fold Cross Validation*. Parameter yang dioptimasi meliputi *criterion*, *max_depth*, *max_features*, *min_samples_leaf*, *min_samples_split*, dan *n_estimators*. Sebanyak 144 kombinasi parameter dievaluasi untuk memperoleh konfigurasi model terbaik. Nilai terbaik yang diperoleh menghasilkan rata-rata *F1-score* sebesar 94,28% pada proses validasi silang. Penggunaan

criterion entropy memungkinkan model memilih atribut berdasarkan nilai informasi yang diperoleh, sedangkan nilai *max_depth*=20 memberikan keseimbangan antara kemampuan model mempelajari pola data dan mengurangi risiko *overfitting*. Jumlah 100 pohon keputusan juga dinilai cukup untuk menghasilkan prediksi yang stabil tanpa meningkatkan kompleksitas komputasi secara berlebihan.

Evaluasi Model Random Forest

Model Random Forest selanjutnya dievaluasi menggunakan data pengujian untuk mengetahui kemampuan generalisasi model terhadap data yang tidak digunakan pada proses pelatihan.

Tabel 8. Hasil Evaluasi Model *Random Forest*

Metrik	Accuracy	Precision	Recall	F1-score
Nilai	92,37%	94,64%	95,12%	94,88%

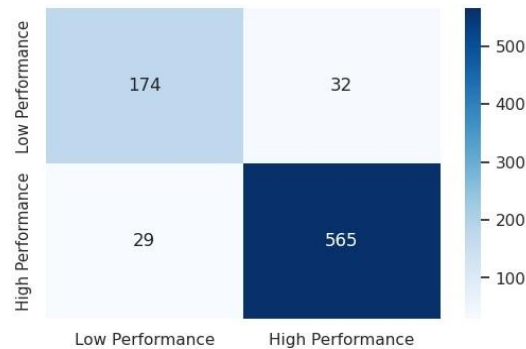
Pada Tabel 8, model *Random Forest* menunjukkan performa klasifikasi yang sangat baik pada data pengujian dengan *accuracy* sebesar 92,37%, yang menandakan sebagian besar video berhasil diklasifikasikan dengan benar ke dalam kategori *High Performance* dan *Low Performance*. Nilai *precision* sebesar 94,64% menunjukkan tingkat ketepatan prediksi yang tinggi, sedangkan *recall* sebesar 95,12% mengindikasikan kemampuan model mengidentifikasi hampir seluruh video pada kelas sebenarnya sehingga *false negative* relatif rendah. Sementara itu, *F1-score* sebesar 94,88% menunjukkan keseimbangan yang baik antara *precision* dan *recall*. Secara keseluruhan, hasil tersebut menunjukkan bahwa *Random Forest* menunjukkan performa klasifikasi yang baik pada dataset penelitian dalam membedakan video *High Performance* dan *Low Performance* berdasarkan metadata yang digunakan.

Tabel 9. Classification Report *Random Forest*

Kelas	Precision	Recall	F1-score	Support
Low Performance	0,86	0,84	0,85	206
High Performance	0,95	0,95	0,95	594
Accuracy			0,92	800
Macro Average	0,90	0,90	0,90	800
Weighted Average	0,92	0,92	0,92	800

Pada Tabel 9, performa model pada setiap kelas menunjukkan bahwa *Random Forest* mampu mengidentifikasi video *High Performance* dengan sangat baik, ditunjukkan oleh nilai *precision*, *recall*, dan *F1-score* masing-masing sebesar 0,95. Sebaliknya, performa pada kelas *Low Performance* sedikit lebih rendah, dengan *precision* sebesar 0,86, *recall* sebesar 0,84, dan *F1-score* sebesar 0,85. Perbedaan ini dipengaruhi oleh ketidakseimbangan distribusi data, di mana jumlah video *High Performance* sebanyak 594 data lebih banyak dibandingkan *Low Performance* hanya 206 data, sehingga model lebih banyak mempelajari karakteristik kelas mayoritas. Meskipun demikian, *macro average* sebesar 0,90 menunjukkan bahwa performa pada kedua kelas tetap berada dalam kategori baik, sedangkan *weighted average* sebesar 0,92 mengindikasikan bahwa model mampu mempertahankan performa klasifikasi yang tinggi dengan mempertimbangkan proporsi masing-masing kelas. Hasil ini menunjukkan bahwa *Random Forest* memiliki kemampuan generalisasi yang baik dan mampu mengklasifikasikan

performa video secara konsisten meskipun terdapat ketidakseimbangan jumlah sampel antar kelas.



Gambar 3. *Confusion Matrix Klasifikasi Random Forest*

Pada Gambar 3, model *Random Forest* berhasil mengklasifikasikan 174 video *Low Performance* dan 565 video *High Performance* dengan benar, sehingga menghasilkan 739 prediksi benar dari 800 data uji, akurasi sekitar 92%. Model juga menghasilkan 32 video *False Positive* pada *Low Performance* diprediksi sebagai *High Performance* dan 29 video *False Negative* pada *High Performance* diprediksi sebagai *Low Performance*. Jumlah kesalahan yang relatif kecil dibandingkan prediksi benar menunjukkan kemampuan model yang baik dalam membedakan kedua kelas berdasarkan metadata video. Selain itu, distribusi *False Positive* dan *False Negative* yang relatif seimbang mengindikasikan bahwa model tidak mengalami bias yang signifikan terhadap salah satu kelas. Hasil ini memperkuat bahwa *Random Forest* memiliki performa klasifikasi yang tinggi dan kemampuan generalisasi yang baik dalam mengidentifikasi video *High Performance* maupun *Low Performance*.

Validasi Menggunakan 10-Fold Cross Validation

Untuk menguji konsistensi model, dilakukan evaluasi menggunakan *10-Fold Cross Validation*. Setiap fold digunakan secara bergantian sebagai data pengujian, sedangkan sembilan *fold* lainnya digunakan sebagai data pelatihan.

Tabel 10. Hasil 10-Fold Cross Validation pada Data Latih

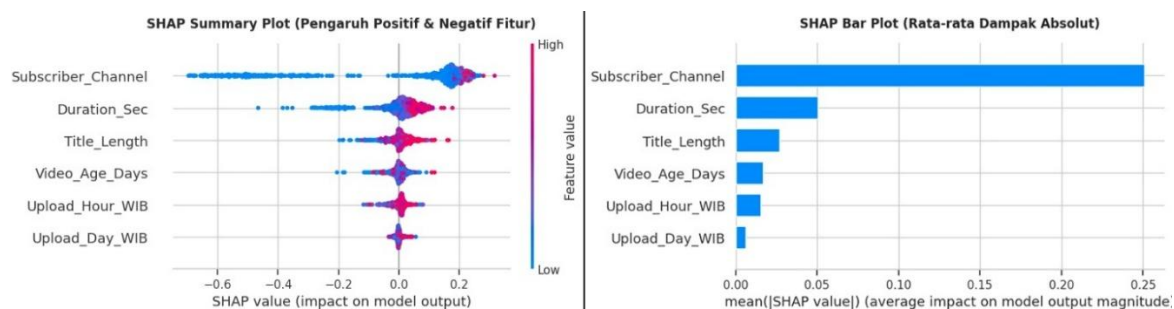
Metrik	Mean (%)	Std Dev (%)
Accuracy	0.9131	0.0163
Precision	0.9420	0.0159
Recall	0.9410	0.0102
F1-score	0.9415	0.0108

Pada Tabel 10, hasil evaluasi *Random Forest* menggunakan *10-Fold Cross Validation* untuk menguji konsistensi performa model pada berbagai pembagian data. Model memperoleh rata-rata *Accuracy* sebesar 91,31% dengan standar deviasi 1,63%, yang menunjukkan performa klasifikasi tinggi dan stabil. Nilai *Precision*, *Recall*, dan *F1-score* masing-masing sebesar 94,20%, 94,10%, dan 94,15%, dengan standar deviasi yang rendah, yaitu 1,59%, 1,02%, dan 1,08%. Rendahnya variasi performa antarfold menunjukkan bahwa model memiliki kemampuan generalisasi yang baik dan tidak bergantung pada pembagian data tertentu. Selain itu, perbedaan yang kecil antara hasil *10-Fold Cross Validation* dan pengujian pada data uji mengindikasikan bahwa model tidak mengalami *overfitting* yang signifikan.

SHAP Explainability

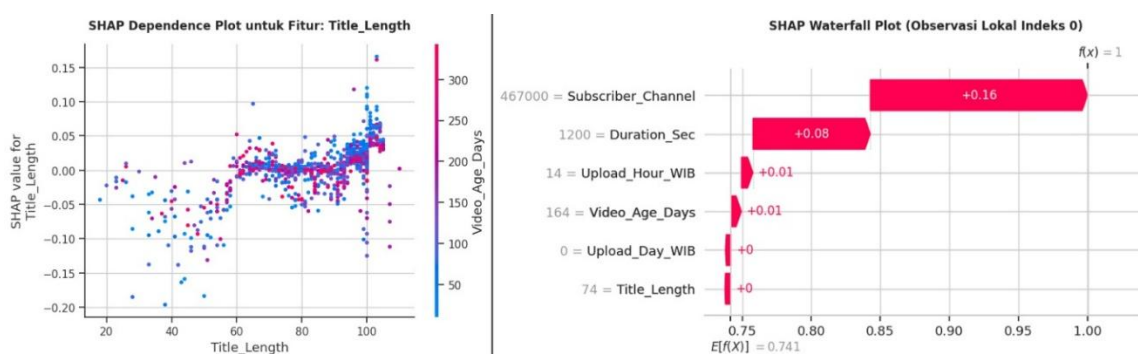
Pembahasan SHAP dalam penelitian ini meliputi SHAP *Summary Plot*, SHAP *Bar Plot*, SHAP *Dependence Plot*, SHAP *Waterfall Plot*, dan SHAP *Force Plot* untuk memperoleh pemahaman

yang lebih komprehensif mengenai faktor-faktor yang memengaruhi klasifikasi performa video Free Fire di YouTube.



Gambar 4. SHAP Summary Plot dan SHAP Bar Plot

Pada Gambar 4, Interpretasi global model *Random Forest* menggunakan SHAP Summary Plot dan SHAP Bar Plot. Pada SHAP Summary Plot, setiap titik merepresentasikan satu video, dengan warna merah menunjukkan nilai fitur tinggi, warna biru nilai rendah, dan posisi pada sumbu horizontal menunjukkan arah pengaruh fitur terhadap prediksi. Nilai SHAP positif meningkatkan peluang video diklasifikasikan sebagai *High Performance*, sedangkan nilai SHAP negatif mengarah pada *Low Performance*. Sementara itu, SHAP Bar Plot menampilkan rata-rata nilai absolut SHAP yang menunjukkan tingkat kepentingan relatif setiap fitur. Kedua visualisasi menunjukkan bahwa *Subscriber_Channel* merupakan fitur paling berpengaruh, diikuti oleh *Duration_Sec*, *Title_Length*, *Video_Age_Days*, *Upload_Hour_WIB*, dan *Upload_Day_WIB*. Nilai *Subscriber_Channel* yang tinggi cenderung menghasilkan SHAP positif sehingga meningkatkan peluang video masuk kategori *High Performance*, sedangkan nilai yang rendah lebih banyak berkontribusi pada *Low Performance*. Pola serupa juga terlihat pada *Duration_Sec* dan *Title_Length*, yang menunjukkan bahwa durasi dan judul yang lebih panjang cenderung memberikan kontribusi positif terhadap performa video. Sebaliknya, *Upload_Hour_WIB* dan *Upload_Day_WIB* memiliki nilai SHAP yang relatif kecil sehingga pengaruhnya terhadap keputusan model lebih rendah. Konsistensi urutan kepentingan fitur pada kedua visualisasi menunjukkan bahwa model *Random Forest* memiliki interpretasi global yang stabil dalam mengidentifikasi faktor-faktor utama yang memengaruhi klasifikasi performa video *Free Fire* di YouTube.

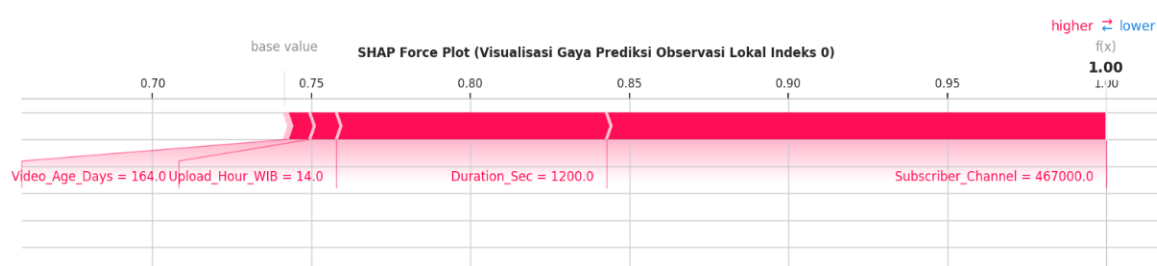


Gambar 5. SHAP Dependence Plot dan SHAP Waterfall Plot

Pada Gambar 5, SHAP Dependence Plot menunjukkan bahwa semakin panjang *Title_Length*, nilai SHAP cenderung meningkat ke arah positif, terutama pada judul dengan panjang di atas sekitar 70 karakter. Hal ini mengindikasikan bahwa judul yang lebih informatif atau deskriptif meningkatkan peluang video diklasifikasikan sebagai *High Performance*, sedangkan judul yang lebih pendek masuk kategori *Low Performance*. Penyebaran titik juga menunjukkan adanya interaksi *Title_Length* dengan *Subscriber_Channel*, yang ditunjukkan oleh gradasi warna, sehingga pengaruh panjang judul tidak berdiri sendiri, tetapi dipengaruhi oleh

ukuran *audiens* kanal. Hasil ini menunjukkan bahwa optimasi judul dapat meningkatkan peluang performa video, meskipun efektivitasnya tetap bergantung pada jumlah *subscriber* yang dimiliki kanal.

SHAP *Waterfall Plot* menjelaskan proses pengambilan keputusan model terhadap satu sampel video, di mana nilai dasar berubah secara bertahap akibat kontribusi setiap fitur hingga menghasilkan prediksi akhir. *Subscriber_Channel* memberikan kontribusi positif terbesar terhadap probabilitas video masuk kategori *High Performance*, diikuti *Duration_Sec* dan *Upload_Hour_WIB* dengan kontribusi positif yang lebih kecil. Sebaliknya, *Title_Length* dan *Video_Age_Days* memberikan kontribusi negatif sehingga sedikit menurunkan probabilitas prediksi. Namun, akumulasi kontribusi positif tetap lebih besar sehingga video diklasifikasikan sebagai *High Performance*. Visualisasi ini menunjukkan bahwa keputusan *Random Forest* merupakan hasil kombinasi kontribusi seluruh fitur, bukan hanya satu variabel.



Gambar 6. SHAP Force Plot

Pada Gambar 6, SHAP Force Plot memvisualisasikan kontribusi masing-masing fitur terhadap prediksi model untuk satu sampel video. Panah merah menunjukkan fitur yang meningkatkan probabilitas prediksi ke kategori *High Performance*, sedangkan panah biru menunjukkan fitur yang menurunkannya. Pada sampel yang dianalisis, *Subscriber_Channel* memberikan kontribusi positif paling dominan sehingga menjadi faktor utama yang mendorong prediksi ke kelas *High Performance*, sedangkan *Title_Length* dan *Video_Age_Days* memberikan kontribusi negatif dengan pengaruh yang lebih kecil. Visualisasi ini menghasilkan interpretasi lokal yang transparan dan mudah dipahami.

SIMPULAN

Penelitian ini mengembangkan pendekatan yang mengintegrasikan pelabelan otomatis berbasis *K-Means*, *Random Forest*, dan *Explainable Artificial Intelligence* (XAI) berbasis SHAP untuk mengklasifikasikan performa video *Free Fire* di *YouTube* berdasarkan metadata video. Pelabelan dilakukan melalui pembentukan dua klaster, yaitu *High Performance* dan *Low Performance*, yang selanjutnya digunakan sebagai target pada proses klasifikasi. Hasil penelitian menunjukkan bahwa model *Random Forest* mampu mengklasifikasikan kedua kelompok performa tersebut dengan akurasi sebesar 92,37% dan rata-rata akurasi 91,31% pada *10-Fold Cross Validation*, yang mengindikasikan bahwa model memiliki performa dan konsistensi yang baik pada dataset penelitian. Selain itu, analisis SHAP menunjukkan bahwa *Subscriber_Channel*, *Duration_Sec*, *Title_Length*, *Video_Age_Days*, *Upload_Hour_WIB*, dan *Upload_Day_WIB* merupakan faktor-faktor yang paling berkontribusi terhadap keputusan model, sehingga memberikan interpretasi yang lebih transparan mengenai pengaruh setiap fitur terhadap hasil klasifikasi. Penelitian ini masih memiliki beberapa keterbatasan, yaitu hanya menggunakan dataset video *Free Fire* dari kanal *YouTube* Indonesia, sehingga hasilnya

belum dapat digeneralisasikan pada kategori konten maupun wilayah lain. Selain itu, label *High Performance* dan *Low Performance* dibentuk secara otomatis menggunakan algoritma *K-Means* berdasarkan karakteristik data, sehingga tidak merepresentasikan label performa aktual yang ditetapkan oleh *YouTube* atau berdasarkan penilaian pakar. Dataset juga dikumpulkan dalam rentang waktu 13 Juni 2025 hingga 14 Juni 2026, sehingga karakteristik performa video berpotensi berubah mengikuti perkembangan algoritma rekomendasi *YouTube* dan perilaku pengguna. Oleh karena itu, penelitian selanjutnya disarankan untuk melakukan validasi label menggunakan acuan eksternal, memperluas cakupan kategori konten dan wilayah penelitian, serta mengintegrasikan fitur multimodal, seperti *Natural Language Processing* (NLP) dan *Computer Vision*, guna meningkatkan kemampuan generalisasi dan interpretabilitas model.

DAFTAR PUSTAKA

- [1] A. Goncalves and Y. M. M. Ng, "Global YouTube Trending Dataset (2022-2025): Three Years of Platform-Curated, Cross-National Trends in Digital Culture," *Proceedings of the Twentieth International AAAI Conference on Web and Social Media*, vol. 20, no. 1, pp. 2817–2827, May 2026, doi: 10.13012/B2IDB-9307654_V1.
- [2] J. Ohme *et al.*, "Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking," *Commun. Methods Meas.*, vol. 18, no. 2, pp. 124–141, 2024, doi: 10.1080/19312458.2023.2181319.
- [3] G. S. Buana, R. Tyasnurita, N. C. Puspita, R. A. Vinarti, and F. Mahananto, "Inter Text-Based Content Analysis on Social Media Using Topic Modelling to Support Digital Marketing," *International Journal On Informatics Visualization*, vol. 1, no. 8, pp. 88–95, Mar. 2024, doi: 10.62527/joiv.8.1.1636.
- [4] F. Ramadani and F. Dafit, "Dampak Game online Free Fire terhadap Karakter Tanggungjawab dan Hasil Belajar Siswa Sekolah Dasar," *Aulad: Journal on Early Childhood*, vol. 3, no. 3, pp. 478–485, Dec. 2023, doi: 10.31004/aulad.v6i3.567.
- [5] C. O. G. A. D. H. Uray, Vika, and C. Cahyaningtias, "Menelaah Kesuksesan Konten YouTube yang Trending Analisis Data Video Trending di YouTube," *Informatik : Jurnal Ilmu Komputer*, vol. 20, no. 2, pp. 84–92, Aug. 2024, doi: <https://doi.org/10.52958/iftk.v20i2.7517>.
- [6] G. Liadeli, F. Sotgiu, and P. W. J. Verlegh, "A Meta-Analysis of the Effects of Brands' Owned Social Media on Social Media Engagement and Sales," *J. Mark.*, vol. 87, no. 3, pp. 406–427, May 2023, doi: 10.1177/00222429221123250.
- [7] W. Zhou, "Predicting Engagement and Dissemination of Digital Cultural and Creative Content Using Deep Learning," *Proceedings of the 2nd International Conference on Digital Management and Information Technology*, vol. 1, no. 1, pp. 263–267, Jun. 2026, doi: 10.1145/3808707.3808748.
- [8] C. Zachlod, O. Samuel, A. Ochsner, and S. Werthmüller, "Analytics of social media data – State of characteristics and application," *J. Bus. Res.*, vol. 144, pp. 1064–1076, May 2022, doi: 10.1016/j.jbusres.2022.02.016.
- [9] M. S. Rahman and H. Reza, "A Systematic Review Towards Big Data Analytics in Social Media," *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 228–244, Sep. 2022, doi: 10.26599/BDMA.2022.9020009.

- [10] M. Suyal and S. Sharma, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *Journal of Artificial Intelligence and Systems*, vol. 6, no. 1, pp. 85–95, Apr. 2024, doi: 10.33969/ais.2024060106.
- [11] M. Annas and S. N. Wahab, "Data Mining Methods: K-Means Clustering Algorithms," *International Journal of Cyber and IT Service Management*, vol. 3, no. 1, pp. 40–47, Mar. 2023, doi: 10.34306/ijcitsm.v3i1.122.
- [12] D. Remawati, H. Wijayanto, Y. R. W. Utami, and B. D. Raharja, "Pengelompokan Film Trending di Youtube Menggunakan TF-IDF dan K-Means Clustering," *Jurnal Sistem Informasi Tgd*, vol. 4, pp. 65–74, Jan. 2025, doi: 10.53513/jursi.v4i1.10614.
- [13] H. Y. A. Putri, P. H. Saputra, R. Y. Doloksaribu, and J. Heikal, "Customer Segmentation Using K-Means Clustering Analysis (Case Study on Amazon Prime Video Userbase)," *Jurnal Cendekia Ilmiah*, vol. 4, no. 4, pp. 963–976, Jan. 2025, doi: 10.56799/jceki.v4i4.8860.
- [14] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Jurnal of Machine Learning*, vol. 2024, no. 1, pp. 69–79, Jun. 2024, doi: <https://doi.org/10.58496/BJML/2024/007>.
- [15] M. H. W. Wahid, Z. Razilu, and M. A. Yasin, "Perbandingan Metode Random Forest Dan Svm Dalam Klasifikasi Sentimen Komentar Youtube Pada Film 'Dirty Vote,'" *Jurnal Ilmiah Penelitian dan Pembelajaran Informatika*, vol. 11, no. 2, pp. 1371–1384, Jun. 2026, doi: 10.29100/jipi.v11i2.7801.
- [16] A. F. Azmi and A. Voutama, "Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Random Forest Dan Decision Tree Dengan Evaluasi Confusion Matrix," *Jurnal Ilmu Komputer dan Informatika*, vol. 13, no. 1, pp. 111–119, Apr. 2024, doi: 10.34010/komputa.v13i1.12639.
- [17] A. M. Salih *et al.*, "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," *Advanced Intelligent Systems*, vol. 7, no. 1, pp. 1–8, Jan. 2025, doi: 10.1002/aisy.202400304.
- [18] A. S. Yazid, Y. Wicaksono, E. Setiawan, and R. Nurhadi, "Implementasi Explainable AI dalam Klasifikasi Kekuatan Password," *Jurnal Sains Manajemen Informatika dan Komputer*, vol. 24, no. 2, pp. 193–202, Aug. 2025, doi: 10.53513/jis.v24i2.12130.
- [19] D. Gaspar, P. Silva, and C. Silva, "Explainable AI for Intrusion Detection Systems: LIME and SHAP Applicability on Multi-Layer Perceptron," *IEEE Access*, vol. 12, no. 1, pp. 30164–30175, Feb. 2024, doi: 10.1109/ACCESS.2024.3368377.
- [20] N. F. Rahman, B. Priyatna, F. Nurapriani, and T. Paryono, "Penerapan Algoritma K-Means Untuk Klasterisasi Film Pada Platform Amazon Prime Video," *Jurnal Mahasiswa Teknik Informatika*, vol. 9, no. 5, pp. 8200–8206, Oct. 2025, doi: <https://doi.org/10.36040/jati.v9i5.15065>.
- [21] A. Sagita, A. Faqih, G. Dwilestari, B. Siswoyo, and D. Pratama, "Penerapan Metode Random Forest Dalam Menganalisis Sentimen Pengguna Aplikasi Capcut Di Google Play Store," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 6, pp. 3307–3313, Dec. 2023, doi: 10.36040/jati.v7i6.8205.
- [22] R. Firdaus, R. Al Hariri, and H. F. Amran, "Sentimen Analisis Masyarakat Tentang Penetapan Hari Raya Idul Adha Tahun 2023 Pada Video Youtube Menggunakan Algoritma Random Forest dan Support Vector Machine," *Jurnal FASILKOM: teknologi inFormASi dan ILmu KOMputer*, vol. 14, no. 1, pp. 278–285, Apr. 2024, doi: 10.37859/jf.v14i1.7012.