Analisis klasifikasi popularitas film Indonesia berdasarkan metadata menggunakan catboost dan SHAP

Aprinia Salsabila Roiqoh, Rizky Parlika*, Firza Prima Aditiawan

Program Studi Informatika, Ilmu Komputer, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Email: 22081010166@student.upnjatim.ac.id, rizkyparlika.if@upnjatim.ac.id, firzaprima.if@upnjatim.ac.id

Abstrak

Popularitas film sering kali dipersepsikan sebagai hasil dari faktor eksternal seperti promosi atau tren pasar, sementara karakteristik internal film yang tercermin dalam metadata pra-rilis belum banyak dianalisis secara sistematis, khususnya dalam konteks perfilman Indonesia. Penelitian ini bertujuan untuk menganalisis karakteristik film populer Indonesia berdasarkan metadata menggunakan pendekatan klasifikasi dan interpretabilitas model. Data yang digunakan berupa metadata film yang mencakup genre, aktor, sutradara, produser, durasi, dan tahun rilis. Popularitas diperlakukan sebagai variabel kategorikal dengan dua kelas, yaitu populer dan tidak populer. Model klasifikasi utama yang digunakan adalah CatBoost, dengan Decision Tree sebagai baseline. Evaluasi dilakukan menggunakan accuracy, precision, recall, F1-score, dan ROC-AUC. Hasil evaluasi menunjukkan bahwa model CatBoost mampu membedakan kedua kelas dengan baik, dengan nilai ROC-AUC sebesar 0,79 dan F1-score sebesar 0,81, yang mengindikasikan bahwa metadata film mengandung informasi yang relevan untuk membedakan film populer dan tidak populer. Untuk memahami kontribusi setiap fitur, digunakan metode SHapley Additive exPlanations (SHAP). Hasil analisis SHAP menunjukkan bahwa sutradara dan produser merupakan fitur dengan pengaruh paling dominan terhadap popularitas film, diikuti oleh durasi film, genre, dan tahun rilis, sementara aktor dan penulis memiliki kontribusi yang relatif lebih kecil secara global. Temuan ini menunjukkan bahwa popularitas film Indonesia lebih dipengaruhi oleh faktor kreatif dan produksi dibandingkan faktor individual semata. Penelitian ini menegaskan bahwa integrasi CatBoost dan SHAP tidak hanya efektif untuk klasifikasi, tetapi juga memberikan pemahaman yang interpretatif mengenai karakteristik film populer di Indonesia.

Kata kunci: popularitas film; metadata; catboost; SHAP; klasifikasi

Indonesian movie popularity classification analysis based-on metadata using CatBoost and SHAP

Abstract

Film popularity is often perceived as the result of external factors such as marketing strategies or market trends, while the internal characteristics reflected in pre-release metadata have received limited systematic analysis, particularly in the context of Indonesian cinema. This study aims to analyze the characteristics of popular Indonesian films based on metadata using a classification and model interpretability approach. The dataset consists of film metadata, including genre, actors, directors, producers, duration, and release year. Popularity is treated as a categorical variable with two classes: popular and non-popular. CatBoost is employed as the main classification model, with Decision Tree used as a baseline. Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The results indicate that CatBoost can effectively distinguish between the two classes, achieving a ROC-AUC of 0.79 and an F1-score of 0.81, which confirms that film metadata contains meaningful information for popularity classification. To interpret the model, SHapley Additive exPlanations (SHAP) is applied. The SHAP analysis reveals that director and producer are the most influential features in determining film popularity, followed by film duration, genre, and release year, while actors and writers contribute relatively less at the global level. These findings suggest that Indonesian film popularity is driven primarily by creative and production-related factors rather than individual elements alone. This study demonstrates that integrating CatBoost and SHAP provides not only accurate classification but also interpretable insights into the characteristics of popular films in Indonesia.

***Keywords**: film popularity; metadata; catboost; SHAP; classification)*

1. Pendahuluan

Industri film telah berkembang menjadi industri bernilai ekonomi tinggi yang sangat kompetitif, sehingga pemahaman terhadap faktor yang memengaruhi popularitas film menjadi kebutuhan strategis. Studi oleh Zhang dan Bai (2023) menunjukkan bahwa industri film kini bergantung pada analisis data historis untuk memahami potensi keberhasilan suatu film, karena proses produksi film melibatkan investasi besar serta risiko kegagalan yang tinggi (Y. Zhang & Bai, 2023). Dalam konteks Indonesia, Sianipar et. al. membuktikan bahwa karakteristik metadata seperti genre mampu menjelaskan perbedaan tingkat popularitas film melalui pendekatan klasifikasi berbasis decision tree (Sianipar et al., 2025). Hal ini menegaskan bahwa metadata pra-rilis merepresentasikan karakter internal film yang dapat dianalisis secara objektif.

Pendekatan berbasis machine learning untuk menganalisis popularitas film berbasis metadata telah banyak dilakukan. Oyewola dan Dada (2022) menunjukkan bahwa atribut seperti genre, aktor, durasi, dan tahun rilis dapat digunakan untuk mengklasifikasikan popularitas film melalui berbagai algoritma klasifikasi, tanpa bergantung pada prediksi pendapatan box office(Oyewola & Dada, 2022). Rantini et al. (2019) mengklasifikasikan film populer dan tidak populer menggunakan Logistic Regression dan SVM dengan fitur genre, komentar, dan likes dari media sosial, yang menunjukkan bahwa pendekatan klasifikasi dua kelas lebih sesuai untuk memahami pola popularitas dibandingkan regresi numerik (Rantini et al., 2019). Zhang dan Bai (2023) juga menemukan bahwa Random Forest memiliki performa terbaik dibandingkan Naive Bayes dan SVR dalam mempelajari pola popularitas berbasis metadata IMDb (Y. Zhang & Bai, 2023). Temuan serupa diperkuat oleh Zhang et al. (2024) yang menunjukkan bahwa kombinasi aktor, sutradara, dan tim produksi memiliki hubungan signifikan dengan popularitas film(Q. Zhang, 2024). Temuan tersebut sejalan dengan Bramantia et al. (2025) yang menunjukkan bahwa kombinasi fitur produksi dan pemeran membentuk pola popularitas film yang berbeda ketika dianalisis menggunakan gradient boosted trees dan teknik clustering(Bramantia et al., 2025). Hasil penelitian tersebut menegaskan bahwa model tree-based ensemble efektif untuk menangkap interaksi kompleks antarfitur metadata film dalam analisis popularitas.

Selain genre dan tim produksi, metadata juga mampu membentuk representasi semantik film. Chen et al. (2023) membuktikan bahwa metadata film dapat dipelajari oleh model untuk membangun representasi adegan, yang menunjukkan bahwa metadata menyimpan struktur informasi yang kaya(Chen et al., 2023). Siddharth et al. (2025) juga mengklasifikasikan genre film menggunakan machine learning, memperlihatkan bahwa metadata dapat memisahkan karakteristik konten secara konsisten (Siddharth et al., 2025)

Namun, sebagian besar model machine learning bersifat black-box, sehingga sulit menjelaskan alasan di balik keputusan klasifikasi. Linardatos et al. (2021) menegaskan bahwa keterbatasan interpretabilitas merupakan hambatan utama dalam penerapan AI di bidang nyata (Linardatos & Papastefanopoulos, 2021). Untuk menjawab masalah tersebut, Lundberg dan Lee (2017) memperkenalkan SHAP yang mampu menjelaskan kontribusi fitur secara adil dan konsisten (Lundberg & Lee, 2017). Filom et al. (2024) memperkuat bahwa SHAP efektif dalam model berbasis pohon karena mampu mengungkap atribusi marjinal fitur (Filom et al., 2024).

Keunggulan SHAP juga dibuktikan secara empiris. Choi dan Abdirayimov (2024) menunjukkan bahwa SHAP dapat menjelaskan karakter dominan dalam klasifikasi karakter Marvel (Choi & Abdirayimov, 2024), sementara Ponce-Bobadilla et al. (2024) memberikan panduan praktis bagaimana SHAP mengungkap pola penting dalam klasifikasi kompleks (Victoria et al., 2024). Dalam konteks industri kreatif, Syamkalla et al. (2024) membuktikan bahwa CatBoost yang dikombinasikan dengan SHAP mampu mengidentifikasi faktor utama popularitas game indie (Syamkalla et al., 2024).

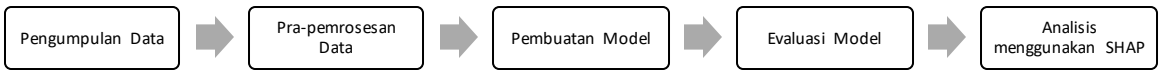
Model berbasis XAI juga mulai diterapkan pada industri film Asia. Leem et al. (2023) mengembangkan model DRECE yang menggabungkan klasifikasi, clustering, dan SHAP untuk menganalisis film Korea secara transparan (Leem et al., 2023). Studi ini menegaskan bahwa integrasi XAI memungkinkan pemahaman yang lebih mendalam terhadap karakteristik film populer, bukan sekadar memprediksi pendapatan.

Berdasarkan seluruh temuan tersebut, penelitian ini bertujuan untuk **menganalisis karakteristik film populer Indonesia** berdasarkan metadata pra-rilis menggunakan **CatBoost** dan **SHAP**, sehingga dapat mengungkap faktor internal yang membedakan film populer dan tidak populer secara interpretatif.

2. Metode

Penelitian ini mengadopsi pendekatan klasifikasi sebagaimana dilakukan oleh Rantini et al. (2019) dan Zhang dan Bai (2023), dengan memperlakukan popularitas sebagai variabel kategorikal dua kelas (Rantini et al., 2019) (Y. Zhang & Bai, 2023). Fitur yang digunakan mengacu pada metadata film yang terbukti relevan dalam penelitian terdahulu, seperti genre, aktor, sutradara, dan produser (Sianipar et al., 2025)(Q. Zhang, 2024). Model utama yang digunakan adalah CatBoost, mengacu pada keberhasilannya dalam menangani fitur kategorikal seperti yang ditunjukkan oleh Syamkalla et al. (2024) dan Leem et al. (2023) . Interpretasi hasil dilakukan menggunakan SHAP, mengikuti pendekatan Lundberg dan Lee (2017) serta panduan praktis dari Ponce-Bobadilla et al. (2024).

Dengan pendekatan ini, penelitian tidak hanya menghasilkan model klasifikasi, tetapi juga memberikan pemahaman menyeluruh mengenai karakteristik metadata yang membentuk popularitas film Indonesia. Alur penelitian secara umum terdiri dari lima tahapan utama, yaitu pengumpulan data, pra-pemrosesan data, pembuatan model, evaluasi model, dan analisis fitur penting, sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Metode Penelitian

Data yang digunakan berupa metadata film Indonesia yang mencakup genre, aktor, sutradara, produser, durasi film, dan tahun rilis. Data diperoleh dari basis data publik perfilman dan dikurasi untuk memastikan kelengkapan serta konsistensi format. Variabel target berupa status popularitas ditentukan berdasarkan ambang batas tertentu dari rating atau jumlah penonton, sesuai dengan ketersediaan data. Seluruh fitur kategorikal dipertahankan dalam bentuk aslinya karena akan diproses langsung oleh algoritma CatBoost.

Tahap pra-pemrosesan meliputi penghapusan data duplikat, penanganan nilai hilang pada variable kategorikal seperti produser maupun variabel numerik seperti tahun rilis. Data kemudian dibagi menjadi data latih dan data uji dengan proporsi 80:20. Tidak dilakukan normalisasi pada fitur kategorikal karena CatBoost memiliki mekanisme internal untuk menangani variabel kategori melalui target statistics.

Model utama yang digunakan adalah CatBoost, yaitu metode gradient boosting berbasis pohon yang dirancang untuk menangani fitur kategorikal secara efisien serta mengurangi bias pada proses pembelajaran (Prokhorenkova et al., 2018). Untuk memberikan pembanding, digunakan pula model Decision Tree sebagai baseline. Pemilihan model ini bertujuan untuk melihat perbedaan kemampuan klasifikasi dan karakteristik interpretasi antar pendekatan berbasis pohon.

Kinerja model dievaluasi menggunakan metrik accuracy, precision, recall, dan F1-score, serta divisualisasikan melalui confusion matrix untuk melihat distribusi kesalahan klasifikasi antara film populer dan tidak populer. Evaluasi ini memastikan bahwa model tidak hanya unggul dari sisi akurasi, tetapi juga seimbang dalam mengklasifikasikan kedua kelas.

Untuk menginterpretasikan hasil klasifikasi, digunakan metode SHapley Additive exPlanations (SHAP) yang diperkenalkan oleh Lundberg dan Lee (2017). Nilai SHAP dihitung untuk setiap observasi dan dirangkum dalam bentuk global feature importance serta dependence plot untuk analisis lokal. Melalui pendekatan ini, kontribusi setiap fitur metadata terhadap klasifikasi film dapat dipahami secara transparan, sehingga karakteristik utama film populer dapat diidentifikasi secara lebih mendalam.

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil analisis dan pembahasan terkait karakteristik film populer di Indonesia berdasarkan metadata film. Model klasifikasi CatBoost digunakan sebagai alat analisis untuk mempelajari pola hubungan antara metadata film dan popularitasnya, yang didefinisikan dalam bentuk kategori populer dan tidak populer. Meskipun evaluasi kinerja model tetap disajikan untuk memastikan validitas analisis, fokus utama bagian ini adalah pada interpretasi kontribusi masing-masing fitur metadata terhadap popularitas film. Untuk tujuan tersebut, metode SHAP digunakan guna mengidentifikasi fitur dan kategori metadata yang memiliki pengaruh paling signifikan, serta untuk memberikan pemahaman yang lebih mendalam mengenai karakteristik film yang cenderung populer.

3.1. Evaluasi Model Klasifikasi

Model CatBoost digunakan untuk melakukan klasifikasi film ke dalam kategori populer dan tidak populer berdasarkan metadata film. Evaluasi model dilakukan untuk memastikan bahwa model memiliki kemampuan diskriminatif yang memadai sebelum digunakan sebagai dasar analisis karakteristik. Perlu ditekankan bahwa evaluasi model dalam penelitian ini tidak bertujuan untuk mencapai tingkat akurasi prediksi maksimum, melainkan untuk memastikan bahwa model dapat menangkap pola yang bermakna dari metadata film. Dengan demikian, hasil evaluasi ini digunakan sebagai landasan untuk analisis interpretabilitas model pada subbagian berikutnya.

Tabel 1. Hasil Evaluasi Model

	Precision	**Recall**	**F1-Score**	**Support**
0	0,82	0,94	0,88	217
1	0,74	0,43	0,54	79
accuracy			0,81	296
Macro avg	0,78	0,69	0,71	296
Weighted avg	0,80	0,81	0,79	296
ROC-AUC			0,79	

Hasil evaluasi menunjukkan bahwa model mencapai akurasi sebesar 0,81 dengan nilai ROC-AUC sebesar 0,795, yang mengindikasikan kemampuan diskriminatif yang cukup baik dalam membedakan kedua kelas. Nilai ROC-AUC yang mendekati 0,8 menunjukkan bahwa model mampu menangkap pola yang relevan dari metadata film tanpa bergantung pada satu fitur tertentu.

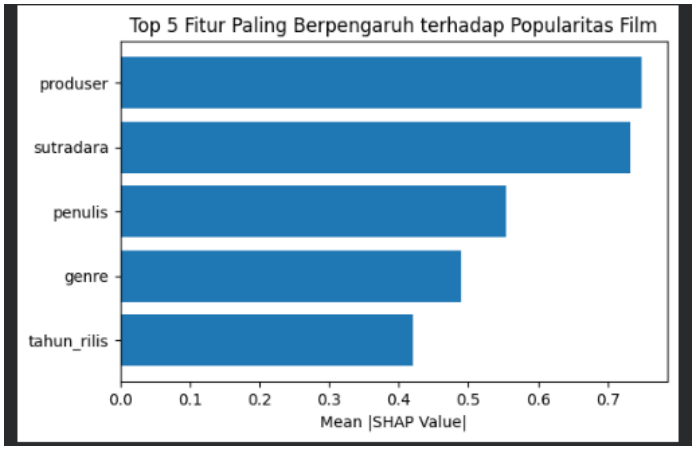
Berdasarkan laporan klasifikasi, kelas tidak populer (kelas 0) memiliki nilai precision sebesar 0,82 dan recall sebesar 0,94, yang menunjukkan bahwa model sangat baik dalam mengidentifikasi film yang tidak tergolong populer. Sebaliknya, untuk kelas populer (kelas 1), model mencapai precision sebesar 0,74 namun recall sebesar 0,43. Hal ini menunjukkan bahwa meskipun sebagian besar film yang diprediksi populer memang benar-benar populer, masih terdapat sejumlah film populer yang tidak berhasil teridentifikasi oleh model.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa model memiliki performa yang cukup andal untuk digunakan sebagai alat analisis karakteristik, sehingga interpretasi fitur menggunakan metode SHAP dapat dilakukan secara valid. Dengan demikian, model ini dinilai memadai untuk mengidentifikasi metadata yang berkontribusi terhadap popularitas film Indonesia.

3.2. Analisis Karakteristik Film Populer Berdasarkan SHAP

Untuk mengidentifikasi metadata yang paling berpengaruh terhadap popularitas film, dilakukan analisis menggunakan metode SHAP (SHapley Additive exPlanations). Analisis SHAP global menunjukkan bahwa produser merupakan fitur dengan kontribusi terbesar terhadap keputusan model, diikuti oleh sutradara dan penulis. Sementara itu, genre dan tahun rilis memiliki pengaruh menengah,

sedangkan aktor dan penulis menunjukkan kontribusi yang relatif lebih kecil secara global. Temuan ini menunjukkan bahwa popularitas film Indonesia lebih dipengaruhi oleh faktor kreatif dan produksi dibandingkan oleh faktor individual semata.



Gambar 2. SHAP Global Feature Importance

Berdasarkan Gambar 2, terlihat bahwa produser dan sutradara merupakan dua fitur metadata dengan kontribusi paling dominan terhadap popularitas film. Tingginya pengaruh produser menunjukkan bahwa aspek produksi memiliki peranan penting dalam menentukan keberhasilan sebuah film, yang mencakup kekuatan manajemen produksi, ketersediaan sumber daya, serta kemampuan dalam mendukung proses distribusi dan promosi. Hal ini mengindikasikan bahwa popularitas film tidak hanya ditentukan oleh konten cerita, tetapi juga oleh bagaimana film tersebut dikelola dan diposisikan dalam industri perfilman.

Selain produser, sutradara juga menunjukkan kontribusi yang sangat signifikan. Temuan ini mengindikasikan bahwa gaya penyutradaraan, reputasi, serta pengalaman kreatif sutradara berperan penting dalam membentuk persepsi kualitas film di mata penonton. Film yang disutradarai oleh individu tertentu cenderung memiliki peluang lebih besar untuk tergolong populer, meskipun pengaruh tersebut tetap bergantung pada kombinasi dengan faktor metadata lainnya.

Fitur penulis dan genre memiliki kontribusi yang berada pada tingkat menengah. Hal ini menunjukkan bahwa kualitas narasi serta jenis cerita yang diangkat tetap berperan dalam menarik minat penonton, namun pengaruhnya bersifat pendukung dan tidak berdiri sendiri. Genre tertentu cenderung lebih sering diasosiasikan dengan film populer, tetapi hasil analisis ini mengindikasikan bahwa genre bukanlah satu-satunya faktor penentu, melainkan berinteraksi dengan faktor kreatif dan produksi lainnya.

Sementara itu, tahun rilis menunjukkan kontribusi yang relatif lebih kecil dibandingkan fitur lainnya. Temuan ini menunjukkan bahwa faktor temporal lebih berfungsi sebagai konteks pendukung yang merepresentasikan kondisi industri perfilman pada periode tertentu, seperti perkembangan jumlah layar bioskop atau tren penonton. Dengan demikian, tahun rilis tidak dapat dianggap sebagai faktor kausal langsung terhadap popularitas film, melainkan sebagai indikator lingkungan industri yang memengaruhi kinerja film secara tidak langsung.

Secara keseluruhan, hasil analisis SHAP menegaskan bahwa popularitas film Indonesia merupakan hasil dari kombinasi faktor produksi dan kreatif, dengan dominasi peran produser dan sutradara, sementara faktor naratif, genre, dan temporal berperan sebagai pendukung. Temuan ini memperkuat pandangan bahwa keberhasilan film tidak ditentukan oleh satu elemen tunggal, melainkan oleh sinergi berbagai metadata yang membentuk karakteristik film secara menyeluruh.

4. Kesimpulan

Penelitian ini bertujuan untuk menganalisis karakteristik film populer Indonesia berdasarkan metadata pra-rilis menggunakan pendekatan klasifikasi dengan CatBoost dan interpretasi model menggunakan SHAP. Hasil evaluasi menunjukkan bahwa model CatBoost mampu mengklasifikasikan

film populer dan tidak populer dengan performa yang memadai, ditunjukkan oleh nilai ROC-AUC sebesar 0,79 dan F1-score sebesar 0,81. Hasil ini mengindikasikan bahwa metadata film mengandung pola yang bermakna untuk membedakan kedua kategori popularitas.

Analisis interpretabilitas menggunakan SHAP menunjukkan bahwa sutradara dan produser merupakan faktor yang paling dominan dalam menentukan popularitas film Indonesia. Temuan ini mengindikasikan bahwa aspek kreatif dan produksi memiliki peranan yang lebih besar dibandingkan faktor individual seperti aktor atau penulis. Selain itu, durasi film, genre, dan tahun rilis berperan sebagai faktor pendukung yang memperkuat karakteristik film populer, meskipun kontribusinya tidak sebesar peran sutradara dan produser.

Secara keseluruhan, penelitian ini menegaskan bahwa popularitas film Indonesia merupakan hasil dari sinergi berbagai metadata yang membentuk karakteristik film secara menyeluruh. Integrasi CatBoost dan SHAP terbukti tidak hanya efektif sebagai alat klasifikasi, tetapi juga sebagai sarana analisis interpretatif untuk memahami faktor internal yang memengaruhi popularitas film. Penelitian ini diharapkan dapat menjadi referensi bagi pelaku industri film dan peneliti dalam mengembangkan strategi produksi dan analisis data di bidang industri kreatif.

Daftar Pustaka

Bramantia, A. C., Hutahaean, J., & Ambarsari, E. W. (2025). *Film Popularity Analysis through Combined K-Means Clustering and Gradient Boosted Trees*. *2*(2), 46–54.

Chen, S., Liu, C.-H., Hao, X., Nie, X., Arap, M., & Hamid, R. (2023). *Movies2Scenes : Using Movie Metadata to Learn Scene Representation*.

Choi, H., & Abdirayimov, S. (2024). *Demonstrating the Power of SHAP Values in AI-Driven Classification of Marvel Characters*. *11*(2), 167–172.

Filom, K., Miroshnikov, A., Kotsiopoulos, K., & Kannan, A. R. (2024). *On marginal feature attributions of tree-based models*. 1–64.

Leem, S., Oh, J., So, D., & Moon, J. (2023). *Towards Data-Driven Decision-Making in the Korean Film Industry : An XAI Model for Box Office Analysis Using*.

Linardatos, P., & Papastefanopoulos, V. (2021). *Explainable AI : A Review of Machine Learning Interpretability Methods*.

Lundberg, S. M., & Lee, S. (2017). *A Unified Approach to Interpreting Model Predictions*. *Section 2*, 1–10.

Oyewola, D. O., & Dada, E. G. (2022). Machine Learning Methods for Predicting the Popularity of Movies. *Journal of Artificial Intelligence and Systems*, *4*(1), 65–82. https://doi.org/10.33969/ais.2022040105

Rantini, D., Inas, R., Purnami, W., Statistika, D., Matematika, F., & Data, S. (2019). *Predicting Popularity of Movie UsingSupport Vector Machines*. *2*(March).

Sianipar, F. D., Irya, A., Syukron, S., Defiyanti, A., Komputer, P. I., Mipa, F., Negeri, U., Jl, M., Iskandar, W., & Estate, M. (2025). Analisis Popularitas Genre Film di Indonesia Menggunakan Algoritma Decision Tree. *Jurnal Mahasiswa Teknik Informatika*, *9*(4), 5555–5563.

Siddharth, A. V, Rakshitha, P., Mazher, S., & Reddy, V. S. (2025). *Movie Genre Classification Using Machine Learning*. *11*(12), 4489–4499.

Syamkalla, M. T., Khomsah, S., Setiya, Y., & Nur, R. (2024). Implementasi Algoritma CatBoost dan SHAPley Additive Explanations (SHAP) dalam Memprediksi Popularitas Game Indie pada Platform STEAM. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(4). https://doi.org/10.25126/jtiik.1148503

Victoria, A., Vanessa, P.-B., Mensing, S., Stodtmann, S., & Maier, C. S. (2024). *Practical guide to SHAP analysis : Explaining supervised machine learning model predictions in drug development Mathematical background*. *October*, 1–15. https://doi.org/10.1111/cts.70056

Zhang, Q. (2024). *Predicting popularity : Machine learning insights into movie team patterns and online ratings*. *25*(3), 386–398.

Zhang, Y., & Bai, Z. (2023). Prediction of movies popularity in supervised learning techniques. *Applied and Computational Engineering*, *29*(1), 142–147. https://doi.org/10.54254/2755-2721/29/20230742