

Klasifikasi *Rating* Film Berdasarkan *Genre* Menggunakan *XGBoost* dan *LightGBM* serta Analisis SHAP

Aprinia Salsabila Roiqoh¹, Rizky Parluka², Firza Prima Aditiawan³

^{1,2,3}Informatika, Ilmu Komputer, Universitas Pembangunan Nasional “Veteran” Jawa Timur

122081010166@student.upnjatim.ac.id, 2rizkyparlika.if@upnjatim.ac.id, 2firzaprima.if@upnjatim.ac.id

Abstract

Movie rating is often used as an indicator of film quality and audience satisfaction. With the large availability of movie data on online platforms, machine learning techniques can be used to analyze the relationship between film characteristics and rating patterns. One important attribute that can influence movie ratings is genre. This study aims to classify movie ratings based on genre using the *XGBoost* and *LightGBM* algorithms and to analyze the contribution of each genre using SHAP (*SHapley Additive Explanations*). Movie data were collected from The Movie Database (TMDB) API and processed through several preprocessing stages including genre separation, data cleaning, one-hot encoding, and rating categorization. The dataset was then divided into training and testing data with a ratio of 70:30. The classification results show that *XGBoost* achieved an accuracy of 0.53, slightly higher than *LightGBM* with an accuracy of 0.52. Further analysis using SHAP indicates that genres such as Horror, Drama, Action, and Comedy have the highest global importance in the classification model. Meanwhile, the analysis of high-rating class predictions shows that Drama has the largest contribution to predicting movies with high ratings. The findings indicate that movie genres have a measurable influence on rating classification, although the importance of genres in the machine learning model does not always align with their average rating values.

Keywords: movie rating classification, genre analysis, *XGBoost*, *LightGBM*, SHAP

Abstrak

Rating film sering digunakan sebagai indikator kualitas film dan tingkat kepuasan penonton. Dengan ketersediaan data film yang semakin besar pada berbagai platform daring, teknik machine learning dapat dimanfaatkan untuk menganalisis hubungan antara karakteristik film dan pola penilaian penonton. Salah satu atribut penting yang dapat mempengaruhi rating film adalah genre. Penelitian ini bertujuan untuk melakukan klasifikasi *rating* film berdasarkan genre menggunakan algoritma *XGBoost* dan *LightGBM* serta menganalisis kontribusi masing-masing genre menggunakan metode SHAP (*SHapley Additive Explanations*). Data film dikumpulkan dari TMDB (*The Movie Database*) API dan diproses melalui beberapa tahap prapemrosesan, yaitu pemisahan genre, pembersihan data, *encoding one-hot*, serta kategorisasi *rating*. Dataset kemudian dibagi menjadi data pelatihan dan data pengujian dengan rasio 70:30. Hasil klasifikasi menunjukkan bahwa model *XGBoost* memperoleh nilai akurasi sebesar 0.53, sedikit lebih tinggi dibandingkan *LightGBM* dengan akurasi sebesar 0.52. Analisis SHAP menunjukkan bahwa genre seperti Horror, Drama, Action, dan Comedy memiliki tingkat kepentingan global yang tinggi dalam model klasifikasi. Sementara itu, analisis pada kelas *rating* tinggi menunjukkan bahwa genre Drama memiliki kontribusi terbesar dalam memprediksi film dengan rating tinggi. Hasil penelitian menunjukkan bahwa genre film memiliki pengaruh terhadap klasifikasi *rating* film, meskipun kontribusi genre dalam model *machine learning* tidak selalu sejalan dengan nilai rata-rata *rating* pada setiap genre.

Kata kunci: klasifikasi *rating* film, analisis genre, *XGBoost*, *LightGBM*, SHAP

©This work is licensed under a Creative Commons Attribution -ShareAlike 4.0 International License

1. Pendahuluan

Industri film merupakan salah satu sektor hiburan global yang terus berkembang dan menghasilkan jumlah data yang sangat besar setiap tahunnya. Data tersebut mencakup berbagai informasi seperti metadata film, ulasan pengguna, serta rating yang diberikan oleh penonton. Platform seperti IMDb menyediakan berbagai informasi terkait film, termasuk genre, aktor, sutradara, serta rating yang diberikan oleh pengguna. Ketersediaan data ini membuka peluang bagi penerapan teknik analisis data dan machine learning untuk memahami pola penilaian film serta faktor-faktor yang mempengaruhi rating film [1,2].

Rating film sering digunakan sebagai indikator penting untuk menilai kualitas suatu film. Film dengan rating yang tinggi biasanya mencerminkan tingkat kepuasan penonton yang lebih baik dibandingkan film dengan rating rendah. Oleh karena itu, analisis terhadap rating film menjadi topik yang banyak diteliti dalam bidang data mining dan machine learning. Berbagai penelitian sebelumnya telah memanfaatkan dataset film untuk berbagai tujuan, seperti sistem rekomendasi film, analisis preferensi penonton, serta prediksi performa film di pasar [3,4].

Salah satu atribut penting dalam analisis film adalah genre. Genre menggambarkan karakteristik utama dari sebuah film dan sering kali berkaitan dengan preferensi penonton terhadap jenis film tertentu. Penelitian

sebelumnya menunjukkan bahwa genre film dapat digunakan untuk menganalisis pola preferensi penonton serta kecenderungan rating yang diberikan oleh pengguna. Penelitian yang menggunakan *dataset MovieLens* menemukan bahwa beberapa *genre* film memiliki rata-rata rating yang lebih tinggi dibandingkan genre lainnya, sehingga genre dapat digunakan sebagai indikator penting dalam memahami pola penilaian penonton terhadap film [5].

Seiring dengan perkembangan machine learning, berbagai algoritma telah digunakan untuk menganalisis data film, termasuk *decision tree*, *support vector machine*, *neural network*, serta metode berbasis *gradient boosting*. Algoritma *gradient boosting* banyak digunakan karena memiliki kemampuan yang baik dalam menangani hubungan *non-linear* pada data serta menghasilkan performa prediksi yang tinggi pada berbagai permasalahan klasifikasi dan regresi [6].

Salah satu algoritma yang banyak digunakan adalah XGBoost (*Extreme Gradient Boosting*). XGBoost merupakan implementasi dari *Gradient Boosting Decision Tree* (GBDT) yang dirancang untuk meningkatkan efisiensi komputasi serta performa prediksi pada *dataset* berukuran besar. Algoritma ini memiliki kemampuan regularisasi yang baik sehingga mampu mengurangi risiko *overfitting* serta meningkatkan akurasi prediksi [7]. Penelitian sebelumnya menunjukkan bahwa model berbasis XGBoost dapat menghasilkan performa prediksi yang baik dalam berbagai permasalahan analisis data, termasuk pada domain kesehatan dan analisis data kompleks [8].

Selain XGBoost, algoritma LightGBM (*Light Gradient Boosting Machine*) juga merupakan metode berbasis *gradient boosting* yang banyak digunakan karena memiliki kecepatan pelatihan yang tinggi serta efisiensi memori yang lebih baik dibandingkan implementasi GBDT konvensional. LightGBM menggunakan teknik *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) untuk meningkatkan efisiensi proses pelatihan model tanpa mengurangi akurasi secara signifikan [9]. Penelitian sebelumnya menunjukkan bahwa LightGBM mampu memberikan performa klasifikasi yang tinggi dan stabil pada berbagai *dataset*, serta memiliki keunggulan dalam efisiensi komputasi dibandingkan metode *boosting* lainnya [10].

Meskipun berbagai penelitian telah memanfaatkan machine learning dalam analisis film, sebagian besar penelitian sebelumnya lebih berfokus pada sistem rekomendasi film atau prediksi popularitas film dengan menggunakan berbagai metadata film seperti aktor, sutradara, anggaran produksi, serta ulasan pengguna [3,11]. Selain itu, terdapat pula penelitian yang menganalisis preferensi *genre* menggunakan pendekatan eksploratif seperti *Exploratory Data Analysis* (EDA), yang hanya berfokus pada distribusi rating dan popularitas *genre* tanpa melibatkan

pemodelan prediktif [5]. Hal ini menunjukkan bahwa penelitian sebelumnya masih terbatas pada analisis deskriptif dan belum mampu memodelkan hubungan antara *genre* film dan *rating* secara langsung.

Di sisi lain, penggunaan algoritma *gradient boosting* modern seperti XGBoost dan LightGBM dalam konteks analisis *genre* film masih relatif terbatas, khususnya untuk tugas klasifikasi rating berbasis *genre* saja (*genre-only features*). Selain itu, sebagian besar penelitian machine learning yang ada hanya berfokus pada performa model tanpa memberikan penjelasan mengenai kontribusi masing-masing fitur terhadap hasil prediksi. Kondisi ini menunjukkan adanya kesenjangan penelitian, yaitu kurangnya integrasi antara model prediktif berbasis *genre* dan pendekatan interpretabilitas model [2].

Untuk mengatasi permasalahan tersebut, pendekatan *Explainable Artificial Intelligence* (XAI) mulai banyak digunakan untuk meningkatkan interpretabilitas model *machine learning*. Salah satu metode interpretasi model yang populer adalah *SHapley Additive Explanations* (SHAP). Metode ini didasarkan pada konsep *Shapley value* dalam teori permainan yang digunakan untuk mengukur kontribusi setiap fitur terhadap hasil prediksi model. Dengan menggunakan SHAP, peneliti dapat mengetahui sejauh mana suatu fitur mempengaruhi hasil prediksi model sehingga model menjadi lebih transparan dan mudah dipahami [12].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membandingkan performa algoritma XGBoost dan LightGBM dalam melakukan klasifikasi rating film berdasarkan *genre*, serta menganalisis kontribusi *genre* terhadap prediksi model menggunakan metode SHAP. Dengan menggabungkan pendekatan klasifikasi dan interpretasi model, penelitian ini diharapkan dapat memberikan pemahaman yang lebih baik mengenai hubungan antara *genre* film dan kategori *rating* film.

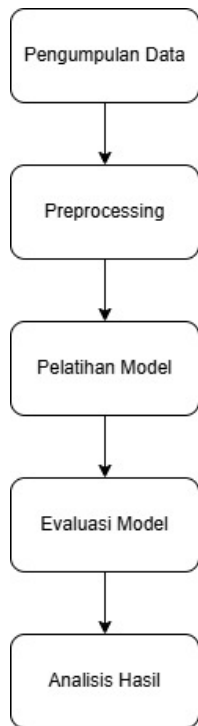
Kontribusi utama dari penelitian ini adalah sebagai berikut:

- Mengimplementasikan algoritma XGBoost dan LightGBM untuk melakukan klasifikasi *rating* film berdasarkan *genre*.
- Membandingkan performa kedua algoritma menggunakan beberapa metrik evaluasi klasifikasi.
- Menggunakan metode SHAP untuk menganalisis kontribusi *genre* terhadap hasil prediksi model sehingga interpretasi model menjadi lebih transparan.

Hasil penelitian ini diharapkan dapat memberikan wawasan mengenai bagaimana *genre* film mempengaruhi klasifikasi *rating* film serta menunjukkan bagaimana teknik *machine learning* dan *explainable AI* dapat digunakan untuk menganalisis data dalam industri film.

2. Metode Penelitian

Penelitian ini menggunakan pendekatan machine learning berbasis klasifikasi untuk menganalisis hubungan antara *genre* film dan *rating* film. Alur penelitian yang dilakukan dalam studi ini dapat dilihat pada Gambar 1.



Gambar 1. Alur Penelitian

Proses penelitian dilakukan melalui beberapa tahapan yang meliputi pengumpulan data, *preprocessing* data, pelatihan model klasifikasi, evaluasi model, serta analisis interpretasi model menggunakan metode SHAP. Pendekatan *machine learning* telah banyak digunakan dalam analisis data film untuk memahami pola preferensi penonton serta memprediksi karakteristik film berdasarkan berbagai atribut yang tersedia [1,2].

2.1. Pengumpulan Data

Data film dikumpulkan menggunakan TMDB (*The Movie Database*) API dengan bahasa pemrograman Python. Proses pengambilan data dilakukan melalui *endpoint discover movie* untuk memperoleh informasi film berupa judul film, *genre*, nilai rating (*vote_average*), serta jumlah suara penonton (*vote_count*).

Film yang diambil merupakan film dengan bahasa asli Inggris (*with original language = en*). Selain itu, dilakukan pembatasan jumlah minimum *vote* sejumlah 50 suara untuk menjaga stabilitas nilai *rating* yang digunakan dalam penelitian. *Dataset* yang dihasilkan berisi 2000 film dengan berbagai kombinasi *genre*.

Genre film yang diperoleh dari TMDB berbentuk daftar *genre ID*. *Genre* tersebut kemudian dikonversi menjadi nama *genre* menggunakan daftar *genre* yang

tersedia pada TMDB API. *Dataset* film yang berisi metadata seperti *genre* dan *rating* banyak digunakan dalam penelitian analisis film karena mampu menggambarkan karakteristik film serta preferensi penonton terhadap film tertentu [13].

2.2. Prapemrosesan Data

Tahap *preprocessing* dilakukan untuk menyiapkan data sebelum digunakan dalam proses klasifikasi. Tahap ini bertujuan untuk meningkatkan kualitas data sehingga model *machine learning* dapat mempelajari pola dalam dataset secara lebih optimal. Proses *preprocessing* yang dilakukan dalam penelitian ini meliputi beberapa tahap sebagai berikut:

a. Pemisahan *Genre*

Film sering kali memiliki lebih dari satu *genre*. Oleh karena itu, *genre* film dipisahkan menjadi beberapa kolom sehingga setiap *genre* dapat diproses secara terpisah dalam tahap *encoding*.

b. Penanganan Nilai Kosong

Beberapa data *genre* dapat memiliki nilai kosong. Pada tahap ini, nilai kosong diisi menggunakan *placeholder* agar data dapat diproses pada tahap selanjutnya.

c. *Encoding Fitur Genre*

Karena *genre film* merupakan data kategorikal, maka data tersebut perlu diubah ke dalam bentuk numerik agar dapat diproses oleh algoritma machine learning. Proses ini dilakukan menggunakan teknik *one-hot encoding*, dimana setiap *genre* direpresentasikan sebagai variabel biner. Teknik *encoding* kategorikal seperti *one-hot encoding* umum digunakan dalam penelitian machine learning untuk mengubah variabel kategorikal menjadi format numerik yang dapat diproses oleh model klasifikasi [14].

d. Kategorisasi *Rating* Film

Nilai *rating* film yang diperoleh dari TMDB berupa nilai numerik dengan skala 0 hingga 10. Dalam penelitian ini, nilai *rating* dikategorikan menjadi empat kelas *rating* untuk membentuk permasalahan klasifikasi [5]. Pembagian Kelas *rating* dapat dilihat pada Tabel 1

Tabel 1. Kategorisasi *Rating* Film

Kelas	Rentang Kelas
0	0.1 – 3.5
1	3.6 – 5.8
2	5.9 – 7.4
3	7.5 - 10

Pembagian rentang kelas dilakukan dengan mempertimbangkan distribusi umum *rating* film yang cenderung berada pada nilai menengah hingga tinggi, sehingga digunakan interval yang tidak seragam untuk merepresentasikan kategori kualitas film secara lebih realistis, yaitu rendah, menengah, tinggi, dan sangat tinggi. Selain itu, pendekatan ini bertujuan untuk menjaga keseimbangan distribusi data antar kelas. Pendekatan kategorisasi *rating* digunakan untuk mempermudah proses klasifikasi serta memungkinkan

model machine learning untuk mempelajari pola hubungan antara *genre* film dan kategori *rating* secara lebih efektif.

2.3. Model Klasifikasi

Setelah tahap *preprocessing* selesai, *dataset* kemudian dipisahkan menjadi data pelatihan dan data pengujian dengan rasio 70:30 menggunakan metode *random split* untuk kemudian di proses menggunakan algoritma. Pada penelitian ini, tidak dilakukan proses hyperparameter tuning, sehingga parameter pada masing-masing algoritma ditentukan secara manual berdasarkan konfigurasi standar yang umum digunakan. Pendekatan ini dipilih untuk menjaga konsistensi perbandingan performa antar model serta memfokuskan analisis pada kemampuan dasar algoritma dalam melakukan klasifikasi. Parameter yang digunakan pada model disajikan pada Tabel 2.

Tabel 2. Parameter Model

Parameter	XGBoost	LightGBM
<i>n_estimators</i>	200	200
<i>max_depth</i>	5	5
<i>learning_rate</i>	0.1	0.1
<i>random_state</i>	42	42

Penelitian ini menggunakan dua algoritma *machine learning* berbasis *gradient boosting*, yaitu *XGBoost* dan *LightGBM*, untuk melakukan klasifikasi kelas rating film berdasarkan fitur *genre*.

2.3.1 XGBoost

XGBoost (*Extreme Gradient Boosting*) merupakan algoritma *machine learning* berbasis *Gradient Boosting Decision Tree* (GBDT) yang banyak digunakan dalam berbagai penelitian karena memiliki performa prediksi yang tinggi serta efisiensi komputasi yang baik. Algoritma ini membangun model secara bertahap dengan menggabungkan beberapa pohon keputusan sehingga menghasilkan model prediksi yang lebih kuat dan akurat.

XGBoost juga memiliki mekanisme regularisasi yang membantu mengurangi risiko *overfitting* serta meningkatkan stabilitas model pada dataset berukuran besar [7].

2.3.2 LightGBM

LightGBM (*Light Gradient Boosting Machine*) merupakan algoritma *machine learning* yang juga berbasis *Gradient Boosting Decision Tree*. Algoritma ini dirancang untuk meningkatkan kecepatan pelatihan model serta efisiensi penggunaan memori.

LightGBM menggunakan dua teknik utama yaitu *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) untuk mengurangi jumlah data dan fitur yang diproses selama pelatihan model tanpa mengurangi akurasi secara signifikan [9].

Dalam penelitian ini, kedua algoritma tersebut digunakan untuk melakukan klasifikasi *rating* film

berdasarkan *genre* dan kemudian dibandingkan performanya.

2.4. Evaluasi Model

Kinerja model klasifikasi dievaluasi menggunakan beberapa metrik evaluasi yang umum digunakan dalam penelitian machine learning, yaitu *Accuracy*, *Precision*, *Recall*, *F1-score*, dan *Confusion Matrix*.

Accuracy digunakan untuk mengukur proporsi prediksi yang benar terhadap seluruh data pengujian. *Precision* dan *recall* digunakan untuk mengevaluasi kemampuan model dalam mengidentifikasi kelas tertentu secara lebih detail, sedangkan *F1-score* digunakan untuk memberikan evaluasi performa model secara lebih seimbang. Penggunaan metrik evaluasi ini banyak digunakan dalam penelitian klasifikasi untuk memberikan evaluasi performa model yang lebih komprehensif [6].

2.5. Analisis SHAP

Untuk memahami kontribusi masing-masing *genre* terhadap hasil prediksi model, penelitian ini menggunakan metode *SHapley Additive Explanations* (SHAP).

SHAP merupakan metode interpretasi model yang didasarkan pada konsep *Shapley value* dalam teori permainan. Metode ini digunakan untuk mengukur kontribusi masing-masing fitur terhadap hasil prediksi model *machine learning*. Dengan menggunakan SHAP, peneliti dapat memahami bagaimana setiap fitur mempengaruhi keputusan model secara lebih transparan [12].

Analisis SHAP dalam penelitian ini dilakukan untuk memperoleh beberapa informasi penting, seperti *Feature importance global*, untuk mengetahui *genre* yang paling berpengaruh terhadap model klasifikasi. Kemudian ada *SHAP summary plot*, untuk melihat distribusi pengaruh setiap *genre* terhadap prediksi model. Dan yang terakhir ada *Kontribusi genre* terhadap kelas *rating* tinggi, untuk mengidentifikasi *genre* yang paling berkontribusi terhadap film dengan rating tinggi. Analisis ini digunakan untuk menjawab pertanyaan penelitian mengenai kontribusi *genre* terhadap *rating* film.

3. Hasil dan Pembahasan

Bagian ini menyajikan hasil eksperimen yang diperoleh dari proses klasifikasi rating film berdasarkan *genre* menggunakan algoritma *XGBoost* dan *LightGBM*. Hasil penelitian meliputi analisis distribusi dataset, rata-rata rating berdasarkan *genre*, perbandingan performa model klasifikasi, serta analisis kontribusi *genre* terhadap prediksi model menggunakan metode *SHAP*.

3.1. Distribusi Dataset

Dataset yang digunakan dalam penelitian ini diperoleh dari *TMDB API* dan terdiri dari film dengan berbagai *genre*. Setelah proses pengumpulan data dan

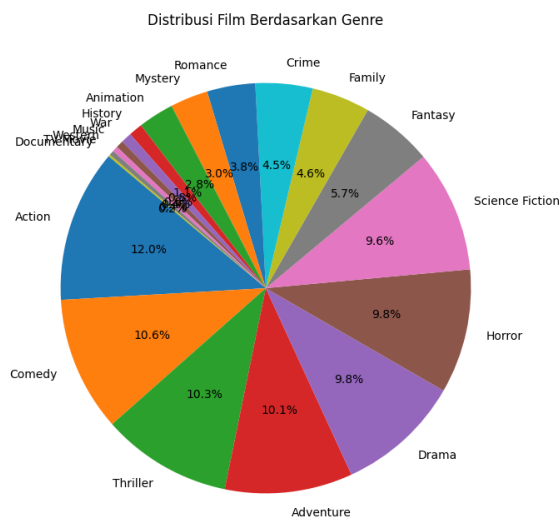
preprocessing dilakukan, diperoleh sejumlah film yang memiliki satu atau lebih *genre*. Untuk memahami karakteristik dataset, dilakukan analisis distribusi jumlah film pada setiap *genre*.

Distribusi jumlah film berdasarkan *genre* ditunjukkan pada Tabel 3.

Tabel 3. Distribusi Jumlah Film Berdasarkan Genre

Genre	Total Film
Action	633
Comedy	559
Thriller	540
Adventure	530
Drama	517
Horror	515
Science Fiction	505
Fantasy	298
Family	242
Crime	237
Romance	200
Mystery	157
Animation	148
History	57
War	44
Music	31
Western	23
TV Movie	20
Documentary	8

Distribusi proporsi film berdasarkan *genre* juga ditunjukkan pada Gambar 2.



Gambar 2. Pie Chart Distribusi Genre Film

Berdasarkan Gambar 2, terlihat bahwa *genre Action* memiliki proporsi terbesar yaitu sekitar 12.0% dari total dataset, diikuti oleh *Comedy* (10.6%), *Thriller* (10.3%), dan *Adventure* (10.1%). *Genre Horror* dan *Science Fiction* juga memiliki proporsi yang cukup besar yaitu sekitar 9–10% dari total *dataset*.

Sebaliknya, beberapa *genre* seperti *Music*, *Western*, *TV Movie*, dan *Documentary* hanya menyumbang proporsi yang sangat kecil terhadap keseluruhan *dataset*. Hal ini

menunjukkan bahwa distribusi *genre* dalam dataset tidak sepenuhnya seimbang.

3.2. Rata-Rata Rating

Untuk memahami hubungan antara *genre* film dan penilaian yang diberikan oleh penonton, dilakukan analisis rata-rata *rating* pada setiap *genre*. Rata-rata *rating* dihitung dengan mengelompokkan film berdasarkan *genre* kemudian menghitung nilai rata-rata *vote average* untuk setiap *genre*. Hasil perhitungan rata-rata *rating* berdasarkan *genre* ditunjukkan pada Tabel 4.

Tabel 4. Rata-Rata Rating

Genre	Rata-Rata Rating
History	7.52
War	7.37
Animation	6.98
Drama	6.86
Crime	6.51
Romance	6.47
Family	6.34
Music	6.33
Adventure	6.21
Western	6.17
Mystery	6.17
Fantasy	5.94
Action	5.86
Comedy	5.79
Thriller	5.71
Science Fiction	5.40
Horror	4.36
TV Movie	3.33
Documentary	3.02

Berdasarkan Tabel 4, terlihat bahwa beberapa *genre* memiliki rata-rata *rating* yang lebih tinggi dibandingkan *genre* lainnya. *Genre History* memiliki rata-rata *rating* tertinggi yaitu sebesar 7.52, diikuti oleh *genre War* (7.37) dan *Animation* (6.98). Hal ini menunjukkan bahwa film dengan *genre* tersebut cenderung memperoleh penilaian yang lebih baik dari penonton.

Genre Drama juga menunjukkan rata-rata *rating* yang relatif tinggi yaitu sebesar 6.86. Hal ini menunjukkan bahwa film drama umumnya memiliki kualitas cerita yang kuat sehingga mampu memperoleh penilaian yang baik dari penonton.

Sebaliknya, beberapa *genre* menunjukkan rata-rata *rating* yang relatif lebih rendah. *Genre Horror* memiliki rata-rata *rating* sebesar 4.36, sedangkan *genre TV Movie* dan *Documentary* memiliki rata-rata *rating* yang paling rendah yaitu masing-masing sebesar 3.33 dan 3.02. Hal ini menunjukkan bahwa film dalam kategori tersebut cenderung memperoleh penilaian yang lebih rendah dibandingkan *genre* lainnya.

3.3. Evaluasi Model

Selanjutnya dilakukan evaluasi performa model klasifikasi yang digunakan dalam penelitian, yaitu *XGBoost* dan *LightGBM*. Kedua model dilatih

menggunakan data pelatihan dan kemudian diuji menggunakan data pengujian untuk memprediksi kelas *rating* film berdasarkan fitur *genre*.

Sebagai pembanding awal, digunakan model baseline berupa *Decision Tree* dengan parameter default tanpa proses *hyperparameter tuning*. Model *baseline* ini digunakan untuk mengetahui sejauh mana peningkatan performa yang dihasilkan oleh algoritma *XGBoost* dan *LightGBM*.

Evaluasi performa model dilakukan menggunakan beberapa metrik evaluasi klasifikasi, yaitu *accuracy*, *precision*, *recall*, dan *F1-score*. Selain itu, analisis juga dilakukan menggunakan *confusion matrix* untuk melihat distribusi prediksi model pada setiap kelas *rating*.

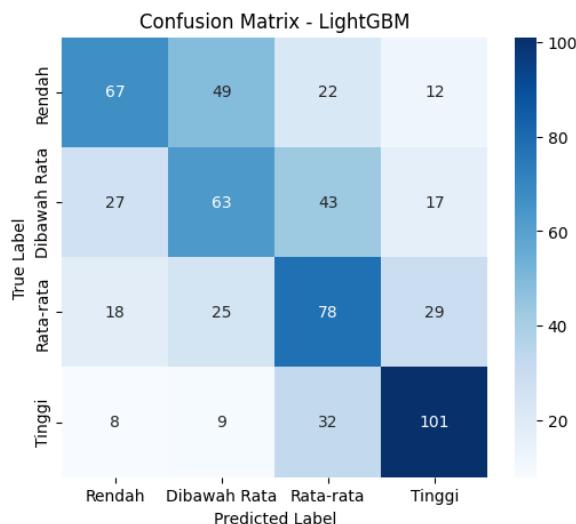
Berdasarkan hasil pengujian, model *baseline Decision Tree* menghasilkan nilai *accuracy* sebesar 0.50, sedangkan model *XGBoost* menghasilkan nilai *accuracy* sebesar 0.53 dan model *LightGBM* sebesar 0.52. Hasil ini menunjukkan bahwa kedua model memiliki performa yang relatif mirip dan hanya sedikit lebih baik dibandingkan *baseline* model, sehingga peningkatan performa yang dihasilkan masih terbatas.

Peningkatan performa yang relatif kecil ini kemungkinan disebabkan oleh keterbatasan fitur yang digunakan, yaitu hanya berbasis *genre* (*genre-only features*). *Genre* film merupakan representasi yang bersifat umum sehingga belum mampu menangkap kompleksitas faktor yang mempengaruhi *rating* film secara menyeluruh, seperti aktor, sutradara, maupun popularitas film.

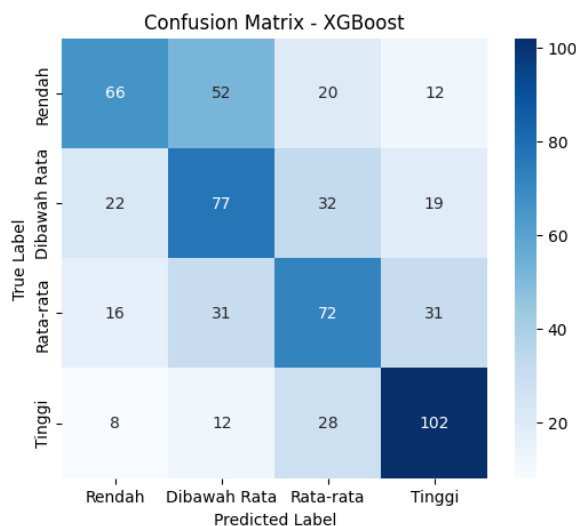
Selain itu, distribusi data antar kelas juga mempengaruhi performa model. Beberapa kelas *rating* memiliki jumlah data yang lebih dominan dibandingkan kelas lainnya, sehingga model cenderung lebih baik dalam memprediksi kelas mayoritas dibandingkan kelas minoritas. Kondisi ini menyebabkan nilai *accuracy* tidak meningkat secara signifikan meskipun model telah menggunakan algoritma yang lebih kompleks.

Selain *accuracy*, nilai *precision*, *recall*, dan *F1-score* pada kedua model juga menunjukkan performa yang cukup seimbang pada setiap kelas *rating*. Pada model *XGBoost*, nilai *F1-score* tertinggi diperoleh pada kelas *rating* tinggi dengan nilai 0.65, sedangkan pada model *LightGBM* nilai *F1-score* tertinggi juga diperoleh pada kelas yang sama dengan nilai yang serupa. Hal ini menunjukkan bahwa kedua model memiliki kemampuan yang cukup baik dalam mengidentifikasi film dengan kategori *rating* tinggi.

Hasil *confusion matrix* untuk masing-masing model ditunjukkan pada Gambar 3 dan Gambar 4.



Gambar 3. Gambar *Confusion Matrix* Model *LGBM*



Gambar 4. Gambar *Confusion Matrix* Model *XGB*

Berdasarkan *confusion matrix* pada Gambar 3 dan Gambar 4, terlihat bahwa sebagian besar prediksi model berada pada diagonal utama matriks, yang menunjukkan bahwa model mampu mengklasifikasikan sebagian besar data dengan benar. Namun demikian, masih terdapat beberapa kesalahan klasifikasi terutama pada kelas *rating* menengah, dimana model terkadang memprediksi kelas tersebut sebagai kelas yang berdekatan.

Kesalahan klasifikasi ini kemungkinan disebabkan oleh kemiripan karakteristik antara film dengan *rating* yang berada pada rentang yang berdekatan. Film dengan *rating* dibawah rata-rata dan rata-rata, misalnya, dapat memiliki karakteristik *genre* yang serupa sehingga model mengalami kesulitan dalam membedakan kedua kelas tersebut.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa algoritma gradient boosting mampu melakukan klasifikasi *rating* film dengan performa yang cukup baik meskipun hanya menggunakan fitur *genre*. Hal ini mengindikasikan bahwa *genre* memiliki pengaruh

terhadap rating film, namun belum cukup kuat untuk menghasilkan model dengan performa tinggi secara mandiri.

3.4 Analisis Interpretasi Model

Untuk memahami kontribusi masing-masing genre terhadap prediksi model klasifikasi, penelitian ini menggunakan metode *SHapley Additive Explanations* (SHAP). Metode ini digunakan untuk menginterpretasikan model machine learning dengan mengukur kontribusi setiap fitur terhadap hasil prediksi model. Dengan menggunakan SHAP, peneliti dapat mengetahui seberapa besar pengaruh setiap genre terhadap proses klasifikasi rating film yang dilakukan oleh model.

Analisis SHAP pertama dilakukan untuk mengidentifikasi *genre* yang berkontribusi terhadap klasifikasi film dengan rating tinggi (*rating class 3*). Hasil analisis ditunjukkan pada Tabel 5.

Tabel 5. Kontribusi Genre Pada Klasifikasi Rating Tinggi

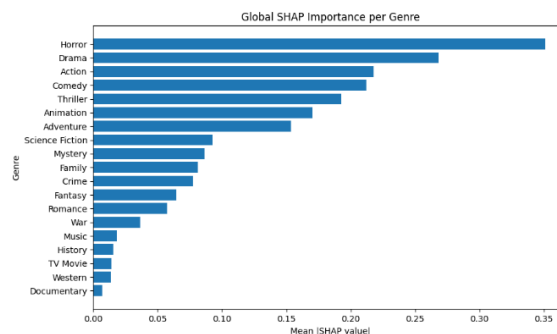
Genre	SHAP Importance	Jumlah Film
Drama	0.588806	295
Horror	0.450593	19
Animation	0.259381	66
Action	0.174649	97
Adventure	0.150123	126
Thriller	0.128779	103
Family	0.109553	67
Science Fiction	0.079763	75
Crime	0.073114	91
Fantasy	0.070060	61

Berdasarkan Tabel 5, terlihat bahwa *genre* Drama memiliki kontribusi terbesar terhadap prediksi kelas *rating* tinggi, dengan nilai SHAP sebesar 0.588806 dan jumlah film sebanyak 295 pada kelas tersebut. Hal ini menunjukkan bahwa keberadaan *genre* drama dalam sebuah film memiliki kecenderungan yang kuat untuk meningkatkan kemungkinan film tersebut diklasifikasikan sebagai film dengan *rating* tinggi oleh model.

Selain itu, *genre* seperti *Horror*, *Animation*, dan *Action* juga menunjukkan kontribusi yang cukup tinggi. Menariknya, *genre Horror* memiliki nilai kontribusi tinggi meskipun jumlah filmnya relatif sedikit, yang mengindikasikan bahwa *genre* ini memiliki pengaruh kuat terhadap keputusan model meskipun tidak dominan secara jumlah data.

Analisis ini menunjukkan bahwa tidak hanya frekuensi kemunculan *genre* yang berpengaruh, tetapi juga bagaimana *genre* tersebut berkontribusi terhadap pola yang dipelajari oleh model.

Selain analisis lokal, dilakukan juga analisis global menggunakan SHAP untuk melihat tingkat kepentingan fitur secara keseluruhan, seperti ditunjukkan pada Gambar 5.



Gambar 5. Global SHAP Importance

Berdasarkan Gambar 5, terlihat bahwa *genre Horror* memiliki tingkat kepentingan global tertinggi, diikuti oleh *genre Drama*, *Action*, dan *Comedy*. Hal ini menunjukkan bahwa *genre-genre* tersebut merupakan fitur yang paling sering digunakan oleh model dalam menentukan kelas rating film. *Genre* lain seperti *Thriller*, *Animation*, dan *Adventure* juga menunjukkan kontribusi yang cukup besar terhadap model klasifikasi.

Sebaliknya, beberapa *genre* seperti *History*, *TV Movie*, *Western*, dan *Documentary* menunjukkan nilai kepentingan yang relatif rendah dalam model. Hal ini menunjukkan bahwa *genre-genre* tersebut memiliki pengaruh yang lebih kecil terhadap proses klasifikasi rating film.

Selain itu, analisis SHAP memberikan pemahaman yang lebih mendalam mengenai bagaimana model melakukan proses pengambilan keputusan. Nilai SHAP tidak hanya menunjukkan tingkat kepentingan suatu *genre*, tetapi juga arah kontribusinya terhadap prediksi model. *Genre* dengan nilai SHAP yang tinggi menunjukkan bahwa fitur tersebut memiliki pengaruh signifikan dalam meningkatkan atau menurunkan kemungkinan suatu film diklasifikasikan ke dalam kelas rating tertentu. Hal ini menunjukkan bahwa model tidak hanya mempertimbangkan keberadaan *genre* secara sederhana, tetapi juga memanfaatkan pola distribusi dan variasi data dalam menentukan keputusan klasifikasi.

Temuan ini mengindikasikan bahwa model machine learning lebih sensitif terhadap dinamika data dibandingkan nilai rata-rata semata, sehingga *genre* dengan distribusi rating yang lebih bervariasi cenderung memberikan kontribusi yang lebih besar terhadap model. Dengan demikian, penggunaan SHAP tidak hanya meningkatkan transparansi model, tetapi juga membantu mengungkap pola tersembunyi dalam hubungan antara *genre* film dan klasifikasi rating. Dengan demikian, analisis SHAP tidak hanya meningkatkan transparansi model, tetapi juga memberikan wawasan bahwa beberapa *genre* memiliki pengaruh yang lebih dominan dalam menentukan *rating* film dibandingkan *genre* lainnya.

3.5 Korelasi Hasil

Untuk memahami hubungan antara hasil analisis statistik dan interpretasi model machine learning, dilakukan analisis korelasi terhadap tiga hasil utama penelitian, yaitu rata-rata rating per *genre*, performa model klasifikasi, dan kontribusi fitur berdasarkan analisis SHAP.

Berdasarkan analisis rata-rata rating per *genre*, beberapa *genre* seperti *History*, *War*, *Animation*, dan *Drama* memiliki nilai rata-rata rating yang relatif lebih tinggi dibandingkan *genre* lainnya. Sebaliknya, *genre* seperti *Horror*, *TV Movie*, dan *Documentary* menunjukkan rata-rata rating yang lebih rendah. Analisis ini menggambarkan kecenderungan umum penilaian penonton terhadap film berdasarkan *genre*.

Namun demikian, hasil analisis global SHAP *importance* menunjukkan pola yang berbeda. *Genre Horror* justru memiliki nilai kepentingan global tertinggi dalam model klasifikasi, diikuti oleh *genre Drama*, *Action*, dan *Comedy*. Perbedaan ini menunjukkan bahwa *genre* dengan rata-rata rating tertinggi tidak selalu menjadi fitur yang paling penting dalam model *machine learning*.

Salah satu faktor yang dapat menjelaskan perbedaan ini adalah distribusi jumlah *film* pada setiap *genre* dalam dataset. Beberapa *genre* dengan rata-rata rating tinggi seperti *History* dan *War* memiliki jumlah film yang relatif lebih sedikit dibandingkan *genre* yang lebih umum seperti *Action*, *Comedy*, atau *Horror*. Karena model machine learning belajar dari pola yang muncul dalam data pelatihan, *genre* dengan jumlah data yang lebih besar cenderung memberikan kontribusi yang lebih besar dalam proses pembelajaran model.

Selain jumlah data, variabilitas rating dalam suatu *genre* juga dapat mempengaruhi tingkat kontribusi fitur dalam model. *Genre* seperti *Horror* cenderung memiliki distribusi rating yang lebih bervariasi dibandingkan beberapa *genre* lain. Variasi ini memberikan informasi yang lebih kuat bagi model dalam membedakan kelas rating yang berbeda, sehingga meningkatkan nilai kontribusi SHAP pada *genre* tersebut.

Analisis SHAP yang difokuskan pada kelas rating tinggi (*rating class* 3) juga menunjukkan hubungan yang sebagian konsisten dengan hasil rata-rata rating. *Genre Drama* memiliki kontribusi terbesar terhadap prediksi kelas rating tinggi, diikuti oleh *genre Animation*, *Action*, dan *Adventure*. *Genre-genre* tersebut juga memiliki nilai rata-rata rating yang relatif lebih baik dibandingkan beberapa *genre* lain, sehingga menunjukkan adanya hubungan antara kecenderungan rating tinggi dan kontribusi fitur dalam model klasifikasi.

4. Kesimpulan

Penelitian ini bertujuan untuk mengklasifikasikan rating film berdasarkan *genre* menggunakan algoritma

XGBoost dan *LightGBM* serta menganalisis kontribusi *genre* terhadap prediksi model menggunakan metode SHAP. Data film diperoleh dari TMDb API dan diproses melalui tahap *preprocessing* yang meliputi pemisahan *genre*, pembersihan data, *encoding* fitur *genre*, serta kategorisasi rating film. Dataset kemudian dibagi menjadi data pelatihan dan data pengujian untuk membangun model klasifikasi.

Hasil evaluasi model menunjukkan bahwa kedua algoritma mampu melakukan klasifikasi rating film dengan performa yang cukup baik meskipun hanya menggunakan fitur *genre* sebagai variabel input. Model *XGBoost* menghasilkan nilai akurasi sebesar 0.53, sedangkan *LightGBM* menghasilkan akurasi sebesar 0.52. Hasil ini menunjukkan bahwa kedua model memiliki performa yang relatif serupa dalam mengklasifikasikan rating film berdasarkan *genre*.

Analisis menggunakan metode SHAP menunjukkan bahwa setiap *genre* memiliki tingkat kontribusi yang berbeda terhadap prediksi model klasifikasi. *Genre Horror*, *Drama*, *Action*, dan *Comedy* memiliki tingkat kepentingan global yang tinggi dalam model. Selain itu, analisis pada kelas rating tinggi menunjukkan bahwa *genre Drama* memiliki kontribusi terbesar terhadap prediksi film dengan rating tinggi.

Berdasarkan hasil penelitian, dapat disimpulkan bahwa *genre* film memiliki pengaruh terhadap klasifikasi rating film, namun tidak cukup kuat untuk menghasilkan performa model yang tinggi apabila digunakan sebagai satu-satunya fitur. Hal ini menunjukkan bahwa faktor lain di luar *genre* juga memiliki peran penting dalam menentukan rating film.

Dari sisi praktis, hasil penelitian ini dapat digunakan sebagai dasar awal dalam memahami preferensi penonton berdasarkan *genre*, serta membantu pengembang sistem rekomendasi atau analisis film dalam mengidentifikasi *genre* yang cenderung berkorelasi dengan rating tertentu. Selain itu, penggunaan metode SHAP memberikan transparansi terhadap model sehingga dapat membantu dalam pengambilan keputusan berbasis data.

Untuk penelitian selanjutnya, disarankan untuk mengembangkan model dengan menambahkan fitur lain seperti aktor, sutradara, popularitas, maupun metadata produksi lainnya, serta mengeksplorasi teknik pemodelan yang lebih kompleks atau pendekatan hybrid untuk meningkatkan performa klasifikasi rating film secara lebih optimal.

Daftar Rujukan

- [1] B. Venkateswarlu, N. Yaswanth, A. M. Kumar, U. Satish, K. Dwijesh, and N. Sunanda, "Cinematic Curator: A Machine Learning Approach to Personalized Movie Recommendations," vol. 15, no. 4, pp. 502–509, 2024.
- [2] W. R. Bristi, Z. Zaman, and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques".
- [3] H. Bhowmick, "Comprehensive Movie Recommendation System".
- [4] A. Singh, A. Rawat, S. Rao, S. Jain, and U. Y. Reddy, "A

- Research Paper on Machine Learning based Movie Recommendation System,” pp. 990–997, 2021.
- [5] R. Desviana and V. Yasin, “Analisis Preferensi Pengguna terhadap Genre Film Menggunakan Eksplorasi Data pada Dataset MovieLens,” *J. Ilm. Ilmu Komput. dan Teknol. Inf.*, vol. 2, no. 2, pp. 1–7, 2025.
- [6] M. K. Najib, A. Irawan, F. N. Salsabilla, and S. Nurdiati, “Performance Comparison of Gradient-based Optimizer for Classification of Movie Genres,” pp. 1–18, 2025, doi: 10.21776/ub.ijma.2025.003.01.1.
- [7] S. Tang, “The box office prediction model based on the optimized XGBoost algorithm in the context of film marketing and distribution,” *PLoS One*, vol. 19, no. 10 October, pp. 1–21, 2024, doi: 10.1371/journal.pone.0309227.
- [8] C. Arafat *et al.*, “Perbandingan Algoritma Random Forest Dan Xgboost Untuk Klasifikasi Penyakit Jantung Berdasarkan Data Medis,” vol. 15, no. 2, pp. 430–435, 2025.
- [9] G. Ke *et al.*, “LightGBM : A Highly Efficient Gradient Boosting Decision Tree,” no. Nips, pp. 1–9, 2017.
- [10] S. Handayani and D. Toresa, “Peningkatan Performa Model Gradient Boosting dalam Klasifikasi Stroke Melalui Optimasi Grid Search,” vol. 14, no. 3, pp. 722–728, 2024.
- [11] A. Sachenko, T. Lendiuk, and K. Lipianina-honcharenko, “Evaluation of ensemble machine learning models for movie recommendation systems,” vol. 8472, 2024.
- [12] A. Victoria, P.-B. Vanessa, S. Mensing, S. Stodtmann, and C. S. Maier, “Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development Mathematical background,” no. October, pp. 1–15, 2024, doi: 10.1111/cts.70056.
- [13] M. Bahraminasr, “IMDb data from Two Generations, from 1979 to 2019; Part one, Dataset Introduction and Preliminary Analysis,” pp. 1–12, 2019.
- [14] H. Wang and H. Zhang, “Movie Genre Preference Prediction Using Machine Learning for Customer-Based Information,” vol. 11, no. 12, pp. 1329–1336, 2017.