KAMPUS AKADEMIK PUBLISING
Jurnal Sains Student Research (JSSR)
Vol.4, No.6 Desember 2026
e-ISSN: 3025-9851, p-ISSN : 3025-986X, Hal 89-103
DOI : https://doi.org/10.61722/jssr.v4i6.12775

PREDIKSI RATING FILM DENGAN PERBANDINGAN ALGORITMA CATBOOST DAN RANDOM FOREST

Muhammad Irmanto[1*], Puspita Nurul Sabrina[2], Herdi Ashaury[3]
[1,2,3] Departemen, Program Studi Teknik Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
*Penulis Korespondensi: muhammadirmanto5@gmail.com

Abstract. *The increasingly rapid development of the film industry has caused the number of films produced and published to continue to rise, resulting in an ever-growing volume of available film data. This condition drives the need for methods capable of utilizing film metadata to predict rating values. This study aims to build a film rating prediction model and compare the performance of the CatBoost and Random Forest algorithms using the TMDB 5000 Movies dataset obtained from Kaggle. The initial dataset consists of 4,803 film records with vote_average as the target variable and eight predictor attributes, namely budget, runtime, release_year, genres, original_language, main_country, main_language, and production_company. The research stages include data quality checking, preprocessing, JSON attribute extraction, feature engineering, categorical data transformation, an 80:20 data split, modeling, hyperparameter tuning, evaluation using MAE, RMSE, and R^2, as well as 5-fold cross-validation. Random Forest uses One-Hot Encoding for the retained categorical features, while CatBoost processes certain categorical features natively. The hold-out results show that CatBoost achieved an MAE of 0.6596, an RMSE of 0.9727, and an R^2 of 0.3725, whereas Random Forest achieved an MAE of 0.6604, an RMSE of 0.9778, and an R^2 of 0.3660. In the 5-fold cross-validation, Random Forest achieved a Mean MAE of 0.6709, a Mean RMSE of 0.9942, and a Mean R^2 of 0.3061, while CatBoost achieved a Mean MAE of 0.6719, a Mean RMSE of 0.9948, and a Mean R^2 of 0.3048. Thus, the two algorithms have relatively comparable performance; CatBoost is slightly better on the hold-out, while Random Forest is slightly better on the cross-validation. Both models were then integrated into a web application using Flask as the machine learning service and CodeIgniter 4 as the interface.*
Keywords: *Film Rating Prediction, CatBoost, Random Forest, Machine Learning, TMDB 5000 Movies.*

Abstrak: Perkembangan industri perfilman yang semakin pesat menyebabkan jumlah film yang diproduksi dan dipublikasikan terus meningkat sehingga volume data film yang tersedia semakin besar. Kondisi tersebut mendorong kebutuhan terhadap metode yang mampu memanfaatkan metadata film untuk memprediksi nilai rating. Penelitian ini bertujuan membangun model prediksi rating film serta membandingkan kinerja algoritma CatBoost dan Random Forest menggunakan dataset TMDB 5000 Movies yang diperoleh dari Kaggle. Dataset awal terdiri atas 4.803 data film dengan vote_average sebagai variabel target dan delapan atribut prediktor, yaitu budget, runtime, release_year, genres, original_language, main_country, main_language, dan production_company. Tahapan penelitian meliputi pemeriksaan kualitas data, preprocessing, ekstraksi atribut JSON, feature engineering, transformasi data kategorikal, pembagian data 80:20, pemodelan, hyperparameter tuning, evaluasi MAE, RMSE, dan R^2, serta validasi 5-fold cross-validation. Random Forest menggunakan One-Hot Encoding untuk fitur kategorikal yang dipertahankan, sedangkan CatBoost memproses fitur kategorikal tertentu secara native. Hasil hold-out menunjukkan CatBoost memperoleh MAE 0,6596, RMSE 0,9727, dan R^2 0,3725, sedangkan Random Forest memperoleh MAE 0,6604, RMSE 0,9778, dan R^2 0,3660. Pada 5-fold cross-validation, Random Forest memperoleh Mean MAE 0,6709, Mean RMSE 0,9942, dan Mean R^2 0,3061, sedangkan CatBoost memperoleh Mean MAE 0,6719, Mean RMSE 0,9948, dan Mean R^2 0,3048. Dengan demikian, kedua algoritma memiliki kinerja yang relatif sebanding; CatBoost sedikit lebih baik pada hold-out, sedangkan Random Forest sedikit lebih baik pada cross-validation. Kedua model kemudian diintegrasikan ke aplikasi web menggunakan Flask sebagai layanan machine learning dan CodeIgniter 4 sebagai antarmuka.

Naskah Masuk:11 September 2026; Revisi: 12 September 2026; Diterima: 13 September 2026; Terbit: 22 September 2026

Kata kunci: Prediksi Rating Film, CatBoost, Random Forest, Machine Learning, TMDB 5000 Movies.

1. LATAR BELAKANG

Perkembangan teknologi digital telah membawa perubahan signifikan pada industri perfilman, mulai dari produksi dan distribusi hingga konsumsi film. Berbagai platform digital memungkinkan masyarakat mengakses film dari berbagai negara dengan lebih mudah. Peningkatan jumlah film tersebut menghasilkan volume metadata yang semakin besar dan menimbulkan kebutuhan terhadap informasi yang dapat membantu memahami karakteristik film dan nilai rating yang diperoleh.

Rating merupakan salah satu indikator yang banyak digunakan untuk menggambarkan penilaian pengguna terhadap suatu film. Nilai tersebut dapat menjadi informasi bagi calon penonton dalam mempertimbangkan film yang akan ditonton dan dapat pula digunakan sebagai bahan analisis bagi pelaku industri. Nilai rating tidak terbentuk secara acak, tetapi dapat berkaitan dengan karakteristik film yang tersedia dalam metadata, seperti budget, runtime, tahun rilis, genre, bahasa, negara produksi, dan perusahaan produksi.

Metadata film dapat dianalisis menggunakan data mining dan machine learning untuk mempelajari hubungan antara karakteristik film dan nilai rating. Data mining digunakan untuk menemukan pola dan hubungan dari data, sedangkan machine learning memungkinkan model mempelajari pola historis dan menghasilkan prediksi pada data baru [1]–[4]. Karena vote_average berupa nilai numerik kontinu, permasalahan ini diperlakukan sebagai supervised learning dengan pendekatan regresi.

Random Forest merupakan algoritma ensemble berbasis bagging yang membangun banyak Decision Tree melalui bootstrap sampling dan pemilihan fitur secara acak. Penggabungan prediksi banyak pohon ditujukan untuk mengurangi variansi dan menghasilkan model yang lebih stabil [7], [11]. CatBoost merupakan algoritma gradient boosting yang membangun model secara bertahap serta memiliki mekanisme untuk menangani fitur kategorikal [13], [14]. Perbedaan mekanisme bagging dan boosting tersebut menjadi dasar pemilihan kedua algoritma dalam penelitian ini.

Penelitian terdahulu menunjukkan bahwa CatBoost dan Random Forest telah digunakan pada berbagai permasalahan prediksi berbasis data tabular. Gomes et al. [16] membandingkan CatBoost, KNN, dan deep learning untuk prediksi rating serial televisi. Putri dan Arianto [17] membandingkan metode bagging dan boosting pada data regresi.

Marpid et al. [18] menggunakan Random Forest Regression untuk prediksi rating pada dataset film TMDB, sedangkan A'la et al. [19] menerapkan CatBoost pada sistem rekomendasi. Perbedaan dataset, fitur, preprocessing, dan skenario evaluasi pada penelitian tersebut menyebabkan hasilnya tidak dapat dibandingkan secara langsung.

Berdasarkan kondisi tersebut, penelitian ini membandingkan CatBoost dan Random Forest menggunakan dataset yang sama, atribut prediktor yang sama, tahapan preprocessing yang dikontrol, serta metrik evaluasi regresi yang sama. Penelitian juga melakukan hyperparameter tuning dan 5-fold cross-validation untuk melihat kestabilan performa. Selain evaluasi model, model diimplementasikan ke dalam aplikasi web agar proses prediksi dapat digunakan secara interaktif.

Tujuan penelitian ini adalah: (1) mengidentifikasi dan menyiapkan kualitas dataset metadata film; (2) menerapkan Random Forest untuk prediksi rating; (3) menerapkan CatBoost untuk prediksi rating; (4) membandingkan kinerja kedua algoritma menggunakan MAE, RMSE, dan R^2; dan (5) mengimplementasikan model ke dalam aplikasi berbasis web.

2. KAJIAN TEORITIS

Data Mining dan Machine Learning

Data mining merupakan proses menemukan pola, hubungan, atau informasi yang bermakna dari kumpulan data melalui tahapan pengolahan dan analisis [1], [2]. Machine learning merupakan pendekatan yang memungkinkan sistem mempelajari pola dari data dan menggunakan pola tersebut untuk menghasilkan prediksi pada data baru [3], [4].

Regresi

Regresi merupakan pendekatan supervised learning untuk memprediksi nilai kontinu. Dalam penelitian ini, vote_average menjadi variabel dependen, sedangkan metadata film menjadi variabel independen. Evaluasi regresi dilakukan menggunakan MAE, RMSE, dan R^2 [5], [6].

Random Forest

Random Forest merupakan ensemble learning berbasis bagging yang membangun sejumlah Decision Tree menggunakan bootstrap sampling dan pemilihan subset fitur secara acak. Pada regresi, prediksi akhir diperoleh dari agregasi atau rata-rata prediksi seluruh pohon [7], [11].

CatBoost

CatBoost merupakan algoritma gradient boosting yang membangun model secara sequential. Setiap tree berikutnya berupaya memperbaiki kesalahan prediksi sebelumnya. CatBoost juga menyediakan mekanisme ordered boosting dan pemrosesan fitur kategorikal yang mendukung penggunaan data tabular dengan atribut numerik dan kategorikal [13], [14].

Hyperparameter Tuning

Hyperparameter merupakan parameter yang ditetapkan sebelum proses pelatihan dan dapat memengaruhi kompleksitas serta kemampuan generalisasi model. Penelitian menggunakan RandomizedSearchCV untuk mencari kombinasi parameter yang menghasilkan skor cross-validation R^2 terbaik pada data latih.

Cross Validation

Cross-validation membagi data menjadi beberapa bagian atau fold untuk melakukan pelatihan dan validasi secara bergantian. Penelitian menggunakan 5-fold cross-validation dengan shuffle=True dan random_state=42 untuk memperoleh rata-rata performa dan melihat variasi hasil antarfold.

Evaluasi Model

MAE menghitung rata-rata nilai absolut selisih antara aktual dan prediksi. RMSE merupakan akar dari rata-rata kuadrat error dan lebih sensitif terhadap error besar. R^2 menunjukkan proporsi variasi target yang dapat dijelaskan model. MAE dan RMSE yang lebih rendah menunjukkan error lebih kecil, sedangkan R^2 yang lebih tinggi menunjukkan kemampuan penjelasan yang lebih besar [15].

3. METODE PENELITIAN

Bagian Penelitian menggunakan pendekatan eksperimen komputasional. Alur penelitian meliputi pengumpulan dataset, pemeriksaan kualitas data, preprocessing dan feature engineering, pembagian dataset, pembentukan model CatBoost dan Random Forest, hyperparameter tuning, evaluasi hold-out, 5-fold cross-validation, analisis residual dan feature importance, serta implementasi sistem.

Dataset

Dataset penelitian adalah TMDB 5000 Movies yang diperoleh dari Kaggle. Dataset awal berjumlah 4.803 data film. Variabel target adalah vote_average dan prediktor terdiri

atas budget, runtime, release_year, genres, original_language, main_country, main_language, dan production_company.

Atribut	Tipe	Peran
budget	Numerik	Prediktor
runtime	Numerik	Prediktor
release_year	Numerik	Prediktor
genres	Kategorikal	Prediktor
original_language	Kategorikal	Prediktor
main_country	Kategorikal	Prediktor
main_language	Kategorikal	Prediktor
production_company	Kategorikal	Prediktor
vote_average	Numerik	Target

Preprocessing dan Feature Engineering

Data diperiksa terhadap missing value dan duplikasi. Atribut JSON seperti genres, production_companies, production_countries, dan spoken_languages diparsing untuk mengambil informasi yang dibutuhkan. release_date diubah menjadi release_year, sedangkan informasi negara, bahasa, dan perusahaan produksi diringkas menjadi main_country, main_language, dan production_company. Genres dan production_company kemudian diproses menggunakan MultiLabelBinarizer.

Untuk fitur kategorikal yang dipertahankan pada model, Random Forest menggunakan One-Hot Encoding, sedangkan CatBoost menggunakan native categorical features. Missing value ditangani sesuai prosedur preprocessing final. Setelah preprocessing diperoleh 4.800 data yang digunakan untuk pemodelan.

Pembagian Dataset

Dataset dibagi menggunakan train_test_split dengan rasio 80:20 dan random_state=42. Sebanyak 3.840 data digunakan sebagai data latih dan 960 data sebagai data uji. Data uji tidak digunakan dalam proses pemilihan hyperparameter.

Pemodelan CatBoost

CatBoostRegressor digunakan untuk membangun model regresi. Model memanfaatkan fitur numerik secara langsung dan fitur kategorikal tertentu melalui parameter cat_features. Model baseline menggunakan iterations=500, learning_rate=0,05, depth=6, loss_function=RMSE, dan random_seed=42.

Pemodelan Random Forest

RandomForestRegressor digunakan sebagai model pembanding. Fitur kategorikal diproses melalui One-Hot Encoding dalam Pipeline sehingga transformasi dipelajari dari

data latih dan diterapkan secara konsisten pada data uji. Model baseline menggunakan n_estimators=300 dan random_state=42.

Hyperparameter Tuning

Hyperparameter tuning menggunakan RandomizedSearchCV dengan 20 kombinasi acak, 5-fold cross-validation, scoring R^2, random_state=42, dan n_jobs=-1. Ruang pencarian CatBoost meliputi iterations, depth, learning_rate, dan l2_leaf_reg. Random Forest diuji pada n_estimators, max_depth, min_samples_split, min_samples_leaf, dan max_features.

Evaluasi Model

Model dievaluasi pada data uji menggunakan MAE, RMSE, dan R^2. Selain hold-out evaluation, model terbaik divalidasi menggunakan 5-fold cross-validation. Analisis residual digunakan untuk melihat pola kesalahan, sedangkan feature importance digunakan untuk interpretasi kontribusi fitur.

4. HASIL DAN PEMBAHASAN

Implementasi Dataset

1. Dataset Penelitian

Dataset awal TMDB 5000 Movies berjumlah 4.803 data film. Dataset memuat metadata film dan vote_average sebagai target. Pemeriksaan awal dilakukan untuk memastikan atribut dapat digunakan dalam proses preprocessing dan pemodelan.

2. Seleksi Atribut

Delapan atribut dipilih sebagai prediktor, yaitu budget, runtime, release_year, genres, original_language, main_country, main_language, dan production_company. vote_average digunakan sebagai target. Atribut yang bersifat identitas atau berpotensi memberikan informasi setelah film dipublikasikan tidak digunakan sebagai prediktor.

Preprocessing Data

Preprocessing dilakukan untuk menghasilkan data yang konsisten dan siap digunakan oleh kedua algoritma.

1. Penanganan Missing Value

Missing value diperiksa menggunakan isnull().sum() dan ditangani sesuai prosedur preprocessing final. Setelah pembersihan, dataset yang digunakan untuk pemodelan berjumlah 4.800 data.

2. Penghapusan Data Duplikat

Duplikasi diperiksa menggunakan duplicated(). Tidak ditemukan data duplikat yang perlu dihapus pada dataset final.

3. Transformasi Data

Struktur JSON pada beberapa atribut diparsing. release_date ditransformasikan menjadi release_year, sedangkan production_countries dan spoken_languages digunakan untuk membentuk main_country dan main_language. Informasi production_companies diringkas menjadi production_company.

4. Encoding Data

Encoding dilakukan sesuai kebutuhan algoritma. Random Forest menggunakan One-Hot Encoding untuk fitur kategorikal yang dipertahankan, sedangkan CatBoost menerima fitur kategorikal tertentu secara native.

5. Seleksi Fitur

Seleksi fitur mempertahankan atribut yang relevan dengan tujuan prediksi rating. vote_average tidak digunakan sebagai prediktor karena merupakan target.

6. Transformasi Data Kategorikal

Genres dan production_company diproses menggunakan MultiLabelBinarizer. Fitur original_language, main_country, dan main_language dipertahankan sebagai fitur kategorikal untuk CatBoost dan ditransformasikan menggunakan One-Hot Encoding pada Random Forest.

7. Hasil Preprocessing

Tahapan	Jumlah
Dataset awal	4.803
Data akhir	4.800
Data latih	3.840
Data uji	960

Hasil preprocessing menghasilkan dataset akhir sebanyak 4.800 data. Dataset tersebut kemudian digunakan pada skenario eksperimen yang sama untuk kedua algoritma.

Pembagian Dataset

Pembagian data menggunakan rasio 80:20 dengan random_state=42. Data latih berjumlah 3.840 dan data uji berjumlah 960. Data uji disimpan untuk evaluasi akhir sehingga tidak digunakan dalam pemilihan hyperparameter.

Implementasi Algoritma CatBoost

1. **Penentuan Variabel Prediktor**

Prediktor CatBoost terdiri atas budget, runtime, release_year, genres, original_language, main_country, main_language, dan production_company dengan vote_average sebagai target.

2. **Pembentukan Data Latih dan Data Uji**

CatBoost menggunakan data latih dan data uji yang sama dengan Random Forest, yaitu 80% dan 20%. Kesamaan pembagian menjaga kondisi evaluasi kedua model.

3. **Penentuan Fitur Kategorikal**

Fitur original_language, main_country, dan main_language diberikan kepada CatBoost sebagai categorical features. Genres dan production_company telah melalui MultiLabelBinarizer sehingga direpresentasikan sebagai fitur hasil transformasi.

4. **Pelatihan Model CatBoost**

CatBoostRegressor dilatih dengan mekanisme gradient boosting. Setelah tuning, parameter terbaik adalah iterations=500, depth=7, learning_rate=0,03, dan l2_leaf_reg=9 dengan Best CV R^2 sebesar 0,2822.

5. **Prediksi Rating Film**

Model hasil pelatihan digunakan untuk menghasilkan estimasi vote_average pada data uji. Prediksi dibandingkan dengan nilai aktual untuk menghitung metrik evaluasi.

6. **Evaluasi Model CatBoost**

Metrik	Nilai
MAE	0,6596
RMSE	0,9727
R^2	0,3725

CatBoost menghasilkan MAE 0,6596, RMSE 0,9727, dan R^2 0,3725 pada hold-out evaluation.

Implementasi Algoritma Random Forest

1. **Transformasi Data Kategorikal**

Fitur kategorikal yang dipertahankan ditransformasikan menggunakan One-Hot Encoding dalam Pipeline. Pipeline menjaga agar encoder dipelajari berdasarkan data pelatihan.

2. **Pembentukan Variabel Prediktor**

Random Forest menggunakan prediktor yang sama dengan CatBoost. Setelah encoding, fitur kategorikal direpresentasikan dalam bentuk numerik.

3. **Pembagian Data Latih dan Data Uji**

Random Forest menggunakan 3.840 data latih dan 960 data uji dengan random_state=42.

4. **Pelatihan Model Random Forest**

RandomForestRegressor membangun sejumlah Decision Tree dengan bootstrap sampling dan random feature selection. Parameter terbaik hasil tuning adalah n_estimators=200, max_depth=None, min_samples_split=10, min_samples_leaf=1, dan max_features=sqrt dengan Best CV R^2 sebesar 0,2879.

5. **Prediksi Rating Film**

Model digunakan untuk memprediksi vote_average pada data uji dan hasilnya dibandingkan dengan nilai aktual.

6. **Evaluasi Model Random Forest**

Metrik	Nilai
MAE	0,6604
RMSE	0,9778
R^2	0,3660

Random Forest menghasilkan MAE 0,6604, RMSE 0,9778, dan R^2 0,3660 pada hold-out evaluation.

Hyperparameter Tuning

Hyperparameter tuning dilakukan pada data latih menggunakan RandomizedSearchCV. Pemilihan parameter menggunakan scoring R^2 dan 5-fold cross-validation.

1. **Hyperparameter Tuning CatBoost**

Parameter terbaik CatBoost adalah iterations=500, depth=7, learning_rate=0,03, dan l2_leaf_reg=9 dengan Best CV R^2 sebesar 0,2822.

2. Hyperparameter Tuning Random Forest

Parameter terbaik Random Forest adalah n_estimators=200, max_depth=None, min_samples_split=10, min_samples_leaf=1, dan max_features=sqrt dengan Best CV R^2 sebesar 0,2879.

3. Evaluasi Baseline Menggunakan DummyRegressor

DummyRegressor digunakan sebagai baseline sederhana dengan strategi mean. Baseline tidak mempelajari hubungan metadata dengan target, sehingga digunakan sebagai pembanding terhadap model machine learning.

Model	MAE	RMSE	R^2
DummyRegressor	0,8280	1,2279	-0,0000
Random Forest	0,6604	0,9778	0,3660
CatBoost	0,6596	0,9727	0,3725

Perbandingan Kinerja Algoritma CatBoost dan Random Forest

Metrik	CatBoost	Random Forest
MAE	0,6596	0,6604
RMSE	0,9727	0,9778
R^2	0,3725	0,3660

Pada hold-out evaluation, CatBoost memperoleh nilai MAE dan RMSE yang sedikit lebih rendah serta R^2 yang sedikit lebih tinggi daripada Random Forest. Selisih MAE sebesar 0,0008, RMSE sebesar 0,0051, dan R^2 sebesar 0,0065 menunjukkan bahwa perbedaan performa kedua algoritma relatif kecil.

Validasi Model Menggunakan Cross Validation

Validasi 5-fold cross-validation dilakukan untuk melihat kestabilan performa model pada beberapa pembagian data.

Model	Mean MAE	Mean RMSE	Mean R^2	Std R^2
CatBoost	**0,6719**	**0,9948**	**0,3048**	**0,0727**
Random Forest	**0,6709**	**0,9942**	**0,3061**	**0,0545**

Pada cross-validation, Random Forest memperoleh Mean MAE 0,6709, Mean RMSE 0,9942, dan Mean R^2 0,3061. CatBoost memperoleh Mean MAE 0,6719, Mean RMSE 0,9948, dan Mean R^2 0,3048. Perbedaannya sangat kecil. Hasil ini juga

menunjukkan bahwa keunggulan CatBoost pada hold-out tidak konsisten pada skenario validasi silang.

Analisis Residual

Residual dihitung sebagai selisih antara nilai prediksi dan aktual. Analisis dilakukan melalui visualisasi Actual vs Predicted, distribusi error, boxplot absolute error, serta identifikasi sepuluh observasi dengan absolute error terbesar. Analisis ini digunakan untuk melihat karakteristik kesalahan prediksi yang tidak sepenuhnya tercermin dari metrik agregat.

Analisis Feature Importance

Feature importance digunakan sebagai alat interpretasi untuk melihat kontribusi relatif fitur terhadap prediksi. Interpretasi tidak digunakan sebagai satu-satunya dasar untuk menentukan algoritma yang lebih baik.

1. **Feature Importance CatBoost**

Feature importance CatBoost dihitung dari kontribusi fitur selama pembentukan model gradient boosting. Nilai importance yang lebih tinggi menunjukkan kontribusi relatif yang lebih besar pada model yang dilatih.

2. **Feature Importance Random Forest**

Feature importance Random Forest diperoleh dari kontribusi penurunan impurity pada seluruh Decision Tree. Karena fitur kategorikal mengalami One-Hot Encoding, interpretasi perlu memperhatikan bahwa satu atribut dapat menghasilkan beberapa fitur turunan.

Implementasi Sistem Prediksi Rating Film

Model diimplementasikan ke dalam aplikasi web menggunakan Flask sebagai layanan machine learning dan CodeIgniter 4 sebagai antarmuka aplikasi.

1. **Arsitektur Sistem**

Pengguna berinteraksi melalui frontend CodeIgniter 4. Metadata film dikirim ke layanan Flask untuk diproses dan menghasilkan prediksi. Hasil prediksi kemudian dikembalikan ke frontend untuk ditampilkan.

2. **Implementasi Backend Flask**

Backend menyediakan endpoint /movies, /movie/{title}, /compare/{title}, /predict, dan /metrics untuk pengambilan data, perbandingan prediksi, prediksi film baru, dan penyajian metrik.

3. Implementasi Frontend CodeIgniter 4

Frontend menyediakan menu Dashboard, Prediction, Metrics, dan About. Komunikasi dengan backend dilakukan melalui request API/AJAX.

4. Halaman Dashboard

Dashboard menampilkan informasi penelitian, jumlah dataset, algoritma yang digunakan, serta ringkasan hasil evaluasi.

5. Halaman Prediction

Halaman Prediction menyediakan prediksi film yang terdapat dalam dataset dan prediksi film baru berdasarkan metadata yang dimasukkan pengguna. Hasil dapat menampilkan prediksi CatBoost dan Random Forest serta nilai aktual untuk film yang tersedia.

6. Halaman Metrics

Halaman Metrics menampilkan MAE, RMSE, dan R^2 untuk membantu pengguna melihat hasil evaluasi kedua algoritma.

7. Pengujian Implementasi

Pengujian mencakup input normal, input kosong, tipe data salah, nilai ekstrem, kategori tidak dikenal, dan komunikasi endpoint API. Hasil pengujian pada skripsi menunjukkan skenario implementasi berjalan sesuai rancangan.

Pembahasan Hasil Penelitian

Hasil hold-out menunjukkan CatBoost sedikit lebih baik daripada Random Forest pada ketiga metrik. Namun, 5-fold cross-validation menghasilkan nilai yang sedikit lebih baik untuk Random Forest. Selisih pada kedua skenario relatif kecil sehingga tidak terdapat keunggulan yang konsisten dari salah satu algoritma pada seluruh skenario evaluasi.

Berikut ini Perbedaan tersebut dapat dipahami dari mekanisme kedua algoritma. Random Forest mengurangi variansi melalui penggabungan banyak pohon yang dibangun dari sampel dan subset fitur berbeda, sedangkan CatBoost membangun model secara bertahap melalui boosting. Selain itu, CatBoost memproses fitur kategorikal tertentu secara native, sementara Random Forest menggunakan One-Hot Encoding.

Pemilihan fitur juga mempertimbangkan risiko data leakage. Atribut yang berpotensi tersedia setelah film dipublikasikan, seperti revenue, popularity, dan

vote_count, tidak digunakan sebagai prediktor karena penelitian berfokus pada metadata yang tersedia sebagai karakteristik film dan menghindari penggunaan informasi target atau informasi pascapublikasi. Dengan demikian, prediksi diarahkan pada hubungan metadata terhadap vote_average tanpa memasukkan vote_average sebagai fitur.

Nilai R^2 yang belum mendekati 1 menunjukkan bahwa metadata yang digunakan belum menjelaskan seluruh variasi rating. Rating dapat dipengaruhi oleh faktor lain yang tidak tersedia dalam dataset. Oleh karena itu, hasil penelitian dipahami sebagai kemampuan prediksi berdasarkan metadata yang digunakan, bukan sebagai model yang menjelaskan seluruh faktor pembentuk rating.

DummyRegressor menghasilkan performa yang lebih rendah daripada kedua model machine learning pada hold-out. Hal ini menunjukkan bahwa pemodelan berbasis metadata memberikan informasi prediktif dibandingkan baseline rata-rata, meskipun masih terdapat error prediksi yang cukup besar.

CatBoost dipilih untuk diimplementasikan pada aplikasi karena memperoleh hasil hold-out terbaik pada eksperimen penelitian. Pemilihan tersebut tidak diartikan sebagai bukti bahwa CatBoost secara konsisten lebih baik, karena hasil cross-validation menunjukkan Random Forest sedikit lebih baik. Kedua hasil evaluasi dipertahankan agar kesimpulan penelitian mencerminkan kondisi eksperimen secara utuh.

5. KESIMPULAN DAN SARAN

Kesimpulan

Penelitian berhasil membangun model prediksi rating film menggunakan CatBoost dan Random Forest pada TMDB 5000 Movies. Dataset awal berjumlah 4.803 data dan setelah preprocessing diperoleh 4.800 data untuk pemodelan.

Pada hold-out evaluation, CatBoost memperoleh MAE 0,6596, RMSE 0,9727, dan R^2 0,3725, sedangkan Random Forest memperoleh MAE 0,6604, RMSE 0,9778, dan R^2 0,3660. CatBoost sedikit lebih baik pada skenario tersebut. Pada 5-fold cross-validation, Random Forest memperoleh Mean MAE 0,6709, Mean RMSE 0,9942, dan Mean R^2 0,3061, sedangkan CatBoost memperoleh Mean MAE 0,6719, Mean RMSE 0,9948, dan Mean R^2 0,3048.

Perbedaan hasil yang kecil dan perubahan arah keunggulan antara hold-out dan cross-validation menunjukkan bahwa kedua algoritma memiliki performa relatif sebanding pada dataset yang digunakan. CatBoost kemudian diintegrasikan ke aplikasi sebagai model yang digunakan berdasarkan hasil hold-out terbaik, sementara Random Forest tetap tersedia sebagai model pembanding.

Aplikasi berhasil mengintegrasikan layanan machine learning Flask dengan frontend CodeIgniter 4 sehingga pengguna dapat melakukan prediksi terhadap film pada dataset maupun film baru serta melihat metrik evaluasi.

Saran

Penelitian selanjutnya dapat menggunakan dataset yang lebih besar atau menggabungkan beberapa sumber data untuk meningkatkan variasi karakteristik film. Metadata tambahan seperti cast, director, keywords, sinopsis, dan ulasan pengguna dapat diteliti untuk melihat pengaruhnya terhadap prediksi rating. Optimasi hyperparameter juga dapat dikembangkan menggunakan Bayesian Optimization atau Optuna, serta penelitian dapat membandingkan model dengan XGBoost, LightGBM, Extra Trees, maupun pendekatan deep learning.

DAFTAR REFERENSI

[1] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
[2] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.
[3] Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.
[4] Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.
[5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
[6] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
[7] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
[8] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann.
[9] Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. Lecture Notes in Computer Science, 1857, 1–15.
[10] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189–1232.
[11] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18–22.
[12] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.
[13] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient Boosting with Categorical Features Support. arXiv preprint arXiv:1810.11363.

[14] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased Boosting with Categorical Features. Advances in Neural Information Processing Systems, 31.

[15] Chai, T., & Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? Geoscientific Model Development, 7(3), 1247–1250.

[16] Gomes, A. L., Vianna, G., Escovedo, T., & Kalinowski, M. (2022). Predicting IMDb Rating of TV Series with Deep Learning: The Case of Arrow. Proceedings of the XVIII Brazilian Symposium on Information Systems (SBSI). https://doi.org/10.1145/3535511.3535520

[17] Putri, E. R., & Arianto, D. B. (2024). Perbandingan Performa Algoritma Metode Bagging dan Boosting pada Prediksi Konsentrasi PM10 di Jakarta Utara. Jurnal Teknologi dan Sistem Informasi (TEKNOSI), 10(1), 72–81. https://doi.org/10.25077/TEKNOSI.v10i1.2024.72-81

[18] Marpid, N. N., Kurniawan, Y. I., & Rahayu, S. P. (2025). Analysis of the Movie Database Film Rating Prediction with Ensemble Learning Using Random Forest Regression Method. Jurnal Teknik Informatika (JUTIF), 6(1). https://doi.org/10.52436/1.jutif.2025.6.1.1563

[19] A'la, F. Y., Hartatik, & Riasti, B. K. (2025). A Hybrid Approach for Recommender Systems Based on Alternating Least Squares and CatBoost. Jurnal Teknik Informatika (JUTIF), 6(4). https://doi.org/10.52436/1.jutif.2025.6.4.5002

[20] Rudro, M. F. A., Bhuiyan, M. S. R., Dojana, A., Afiah, Shakib, M. R. H., & Oishi, A. A. (2025). Meta-Ensemble Learning for IMDb Ratings: A Stacked Hybrid Model Integrating Gradient Boosting and Deep Neural Networks. Preprint. https://doi.org/10.22541/au.174223552.29403106/v1

[21] 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