Jurnal Teknik Informatika (JUTIF)
Vol. 6, No. 1, February 2025, pp. 1-10

DOI: https://doi.org/10.52436/1.jutif.2025.6.1.1563
p-ISSN: 2723-3863
e-ISSN: 2723-3871

ANALYSIS OF THE MOVIE DATABASE FILM RATING PREDICTION WITH ENSEMBLE LEARNING USING RANDOM FOREST REGRESSION METHOD

Nuravifah Novembriana Marpid[1], Yogiek Indra Kurniawan*[2], Swahesti Puspita Rahayu[3]

[1,2,3]Informatics, Engineering Faculty, Universitas Jenderal Soedirman, Indonesia
Email: [1]nuravifah.marpid@mhs.unsoed.ac.id, [2]yogiek@unsoed.ac.id, [3]swahesti.rahayu@unsoed.ac.id

(Article received: December 4, 2024; Revision: January 7, 2025; published: February 20, 2025)

Abstract

The film industry has become a very profitable industry. However, during COVID-19 the film industry experienced an unfavorable impact with the delay in the screening schedule of new films, many cinemas were prohibited from operating so they were completely closed, and it wasn't easy to obtain permits to carry out the filmmaking process. To survive in this industry from the impact of the pandemic, it is necessary to consider several factors such as targeted promotion methods by using the right selection of predictive decisions with market and trends. Predicting the success of a film is very helpful in determining the success rating and quality of the film to be released. The Random Forest Regression method is used to conduct predictive analysis on films. This study uses the M-estimate encoding technique to handle categorical data into numerical data, and the result shows that the application of M-estimate encoding increases the correlation value between features. In the Random Forest Regression method with 1000 trees, dividing 80% training data and 20% testing data, the R2 performance score was 86%, the MSE score was 12%, the RMSE score was 35% and the MAE score was 22%. The 10-fold cross-validation score in this study was 85%. This shows that the Random Forest Regression method using 80% training data produces the best performance score.

Keywords: *10-fold cross-validation, Film Rating, M-estimate encoding, Random Forest Regression, TMDB*

ANALISIS PREDIKSI *THE MOVIE DATABASE RATING* FILM DENGAN MENGGUNAKAN *ENSEMBLE LEARNING* MENGGUNAKAN METODE *RANDOM FOREST REGRESSION*

Abstrak

Industri film sudah menjadi industri yang sangat menguntungkan. Namun pada masa *COVID-19*, industri film mengalami dampak yang tidak menguntungkan dengan tertahannya jadwal penayangan film-film baru, banyaknya bioskop yang dilarang untuk beroperasi sehingga tutup total hingga sulitnya izin untuk melakukan proses pembuatan film. Untuk bertahan dalam industri ini dari dampak pandemi, diperlukan pertimbangan beberapa faktor seperti cara promosi yang tepat sasaran dengan menggunakan pemilihan keputusan prediksi yang tepat dengan perkembangan pada pasar dan tren yang sedang terjadi. Prediksi keberhasilan film sangat membantu untuk mengetahui peringkat keberhasilan dan kualitas film yang akan dirilis. Dalam melakukan analisis prediksi pada film digunakan metode *Random Forest Regression*. Penelitian ini memanfaatkan teknik *M-estimate encoding* untuk menangani *categorical data* menjadi *numerical data*, dan didapatkan hasil yang menunjukkan bahwa penerapan *M-estimate encoding* meningkatkan nilai korelasi antar *features*. Pada metode *Random Forest Regression* dengan 1000 pohon pembagian 80% *training data* dan 20% *testing data* didapatkan nilai performa R2 86%, nilai MSE sebesar 12%, nilai RMSE sebesar 35% dan nilai MAE 22%. Untuk nilai *10-fold cross validation* pada penelitian ini didapatkan nilai sebesar 85%. Ini menunjukkan bahwa penggunaan metode *Random Forest Regression* dengan menggunakan 80% *training data* menghasilkan nilai performa terbaik.

Kata kunci: *10-fold cross-validation, M-estimate encoding, Random Forest Regression, Rating* Film, *TMDB*

1. PENDAHULUAN

Industri film sudah menjadi bentuk usaha media yang sangat menguntungkan. Pada pasar global, industri ini mendapatkan keuntungan sampai 101 miliar USD pada tahun 2019 dengan kenaikan sebesar 8% dari tahun sebelumnya. Untuk pasar U.S saja mencapai 36.6 miliar USD dengan kenaikan sebesar 4% dari tahun sebelumnya dan kenaikan

sebesar 25% dari tahun 2015 [1]. Masa pandemi *COVID-19*, industri film mengalami dampak buruk. Banyak film harus ditahan jadwal penayangannya, banyak bioskop harus tutup untuk sementara waktu dan tertundanya proses pembuatan film. Dijelaskan pada tahun 2020 di Indonesia laju pertumbuhan layar dan bioskop mengalami pelambatan. Dilaporkan bahwa jumlah layar hanya naik 1,7% menjadi 2.145 layar dan bioskop bertambah 1,8% menjadi 517 bioskop [2].

Dilansir pada laman dataindonesia, penayangan film KKN Desa Penari pada tahun 2022 mencapai 9.233.847 penonton melampaui film Warkop DKI *Reborn*: Jangkrik *Boss*! *Part* 1 yang ditayangkan pada tahun 2016. Tidak hanya itu, film Pengabdi Setan 2: *Communion* yang ditayangkan pada Juli 2022 menjadi urutan ke-2 dengan jumlah penonton terbanyak dengan jumlah penonton 6.390.970 dan film *Miracle in Cell No.7* meraih urutan ke-3 dengan jumlah penonton sebanyak 5.852.916 [3].

Oleh sebab itu, industri film menjadi industri yang sangat besar [4]. Pengetahuan sebelumnya tentang keberhasilan atau kegagalan tertentu film dan faktor apa yang mempengaruhi kesuksesan akan menguntungkan rumah produksi dalam cara mempromosikan dan keputusan bisnis mahal lainnya. Salah satu cara keberhasilan sebuah film diidentifikasi adalah melalui peringkatnya, memprediksi peringkat film sebelum dirilis berdasarkan data yang tersedia pada saat itu akan membantu dalam pengambilan keputusan [5]. Pakar industri film sepakat bahwa peringkat adalah faktor kunci dalam kesuksesan film dan membantu perusahaan produksi film dan investor untuk mendapatkan kesuksesan finansial. Perusahaan dapat melihat film mana yang cenderung memiliki peringkat yang baik dan membuat strategi memanfaatkan film untuk meningkatkan keuntungan [6].

Prediksi keberhasilan film penting karena melibatkan waktu dan investasi yang signifikan. Jadi penting untuk memiliki banyak akurasi dalam prediksi [7]. Sebuah rumah produksi bisa melakukan penyesuaian dalam merilis filmnya, jika mengetahui lebih dulu kemungkinan keberhasilan terkait film tersebut. Tujuannya untuk mendapatkan keuntungan maksimal setelah film dirilis. Bahkan bisa menggunakan prediksi untuk tahu perkembangan pasar. Ini menunjukkan adanya kebutuhan mendesak pada perangkat lunak terkait untuk dikembangkan [8]. Kualitas atau kesuksesan dari film dapat dinilai dengan menggunakan sistem angka yang disebut *rating. Rating* biasanya diasosiasikan dengan rentang angka 0 sampai 5 atau 0 sampai 10 [9]. *Rating* merupakan suatu penilaian yang diberikan oleh *user* [10]. *Rating* film dipengaruhi oleh banyak faktor, sehingga akurasi dalam memprediksi *rating* sebuah film baru menjadi sebuah tantangan [8]. Dipengaruhi oleh banyak faktor seperti alur cerita, sinematografi, pemeran, musik dan lainnya. Salah satu contoh atribut yang mungkin berpengaruh terhadap kualitas film adalah sutradara dan aktor-aktris. Beberapa sutradara dan aktor-aktris biasanya hanya akan memilih proyek film dengan *screenplay* yang berkualitas. Dengan demikian, sutradara dan aktor-aktris tersebut cenderung memiliki reputasi tinggi karena film yang disutradarai dan dibintangi selalu memiliki *rating* tinggi. Terdapat berbagai *platform online* yang memiliki catatan mengenai informasi detail dari sebuah film [9]. Contohnya adalah *The Movie Database* (TMDB). Dilansir dari *website* resmi *themoviedb*, *The Movie Database* (TMDB) adalah basis data film dan tv yang dibuat oleh komunitas. Setiap bagian data telah ditambahkan oleh komunitas sejak tahun 2008. Rata-rata lebih dari 1.000 gambar ditambahkan setiap hari. TMDB secara resmi mendukung 39 bahasa dan digunakan di lebih dari 180 negara [11].

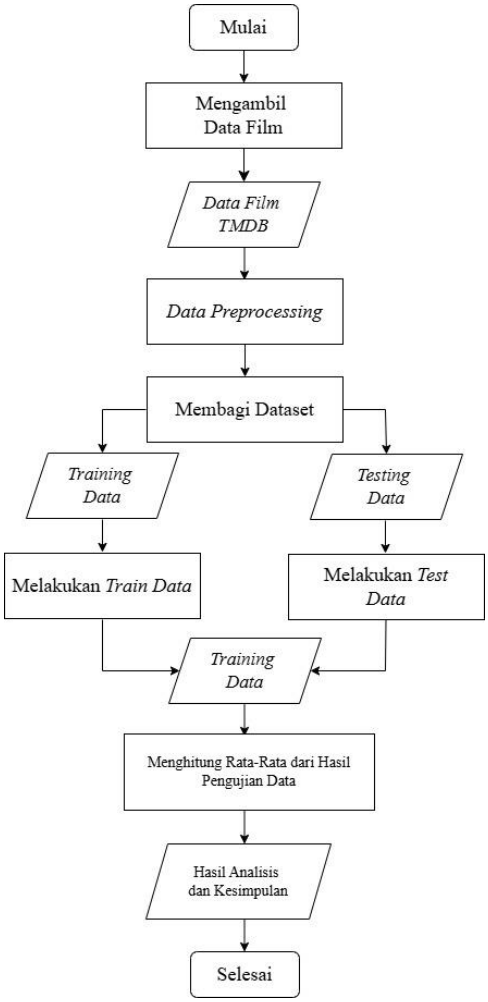
Data mining merupakan salah satu bidang kajian yang dapat menemukan informasi berharga dari sejumlah data yang besar. Terdapat beberapa metode dalam *data mining* yang dapat digunakan untuk menggali informasi berharga, salah satunya prediksi. Prediksi merupakan salah satu metode untuk memperkirakan sebuah nilai pada masa yang akan datang berdasarkan model klasifikasi yang diperoleh sebelumnya. Metode prediksi dalam *data mining* sudah banyak diimplementasikan untuk menyelesaikan beberapa permasalahan, salah satunya adalah prediksi *rating* film [12]. *Data mining*, diharapkan dapat menjadi acuan dan juga penyelesaian masalah terkait prediksi *rating* film.

Metode *Random Forest Regression* diartikan sebagai gabungan dari *Decision Tree.* Banyaknya *Decision Tree* akan mempengaruhi akurasi *Random Forest* secara keseluruhan. Setelah pembentukan *Decision Tree* maka *Random Forest* dapat melakukan klasifikasi. *Random Forest* adalah klasifikasi yang terdiri dari beberapa pohon keputusan [13]. Pada metode *Random Forest*, data maupun atribut diambil secara acak sehingga menghasilkan berbagai model pohon keputusan [10]. *Random Forest* telah banyak digunakan baik untuk klasifikasi dan regresi karena kinerjanya yang unggul dan strukturnya yang sederhana [14]. Kelebihan yang diperkenalkan Breiman dari *Random Forest* diantaranya mampu menghasilkan *error* yang relatif rendah, performa yang baik dalam klasifikasi, dapat mengatasi data pelatihan dalam jumlah besar secara efisien, serta metode yang efektif untuk mengestimasi *missing data* [15].

Berdasarkan latar belakang tersebut penelitian ini bertujuan untuk menganalisis *rating film* dengan data dari TMDB menggunakan metode *Random Forest Regression*, diharapkan dapat menemukan nilai performa yang paling efektif dengan serangkaian teknik *data processing* dan teknik *data manipulation* yang tepat. Sehingga mendapatkan prediksi *rating film* paling akurat demi terpenuhinya tujuan untuk mencapai keberhasilan film. Penelitian

ini juga bertujuan untuk dapat memprediksi *rating* suatu film berdasarkan atribut-atribut yang ada. Pemilihan kombinasi atribut film yang sesuai sebelum memulai produksi dapat meningkatkan kualitas film nantinya [9]. Selain itu, bertujuan untuk meningkatkan pemahaman tentang penerapan tenik dan metode dalam konteks *rating* film. Oleh karena itu, penelitian ini akan memberikan manfaat bagi produsen industri film dan peneliti yang tertarik dalam analisis *rating* film.

2. METODE PENELITIAN



Gambar 1. Mekanisme Penelitian

Mekanisme Penelitian dapat dilihat pada Gambar 1. Metode penelitian ini meliputi tahapan yang terstruktur, dimulai dari pengambilan data film pada *website Kaggle*, persiapan data, pembagian *dataset training* dan *testing*, pemodelan, hingga menemukan hasil analisis dan kesimpulan. Dengan pendekatan ini, memastikan bahwa setiap langkah dalam proses analisis prediksi *rating* film dikembangkan secara sistematis dan terkoordinasi untuk mencapai hasil yang akurat dan relevan dalam penelitian ini.

2.1. Pengambilan Data

Pada langkah awal pengambilan data film TMDB pada *website Kaggle* dilakukan dengan mengunduh *file* berekstensi *Comma Separated Values* (CSV). Dengan demikian, data yang diperoleh mencakup berbagai *id, imdb id, budget, revenue, original title, cast, homepage, director, tagline, runtime, genres, production companies, release date, release year, vote count, vote average* yang merupakan *rating* dari film dari pengguna TMDB.

2.2. *Data Preprocessing*

Data preprocessing ialah proses yang dilakukan diawal yang bertujuan untuk membantu dalam membuat data mentah menjadi data yang berkualitas dalam arti mencegah kesalahan dalam *data mining* [8]. Berdasarkan data film TMDB yang sudah diunduh pada langkah sebelumnya, penulis akan melakukan langkah pembersihan pada data film. Tahapan pada *data preprocessing* dimulai dari *data cleaning* dan *data transformation.*

a. *Data Cleaning*

Tahapan *data cleaning* akan dilakukan penghapusan data yang tidak lengkap dan tidak relevan, penanganan *missing values*, penghapusan terhadap data yang duplikat dan penyesuaian terhadap kesalahan dalam ejaan dan melakukan *filter outlier* pada atribut.

b. *Data Transformation*

Data Transformation merujuk pada proses *generalization* dan *normalization.* Pada penelitian akan dilakukan penggabungan nilai dari atribut pada masing-masing kategori dan menormalisasikan menjadi *numerical data.* Terjadi pengubahan *categorical data* menjadi *numerical data* pada beberapa atribut yang bersifat *categorical.* Seperti pada atribut *genres, director, cast* dan *production companies.* Dengan menggunakan teknik *M-estimate encoding.*

2.3. Pembagian *Dataset*

Setelah data siap digunakan untuk analisis. Maka akan dilakukan pembagian menjadi 2 jenis *dataset. Training dataset* dibentuk untuk melatih model. *Testing dataset* dibentuk untuk mengevaluasi performa dari model. Pembagian akan dibentuk menjadi 80:20. 80% untuk *training dataset. 20% untuk testing dataset.* Pembagian akan dilakukan secara acak. Sehingga *training dataset dan testing dataset* akan mengandung data yang beragam nilainya.

2.4.1. Melakukan *Train Data*

Pada tahap ini *training dataset* akan dipersiapan untuk digunakan dalam model *Random Forest Regression* sebagai bahan pembelajaran untuk analisis prediksi *rating* film. Sehingga nantinya siap untuk memprediksikan *rating* film di masa yang akan datang.

2.4.2. Melakukan *Test Data*

Pada *testing dataset* akan dipersiapkan untuk digunakan dalam *Random Forest Regression* sebagai bahan evaluasi prediksi *rating* film. Untuk mengetahui seberapa akurat dan relevan nantinya pemodelan dibuat dalam penelitian ini.

2.5. Model

Penelitian ini akan menggunakan *Random Forest Regression* dalam pendekatannya. *Random Forest* adalah metode *ensemble learning* yang pertama kali diusulkan oleh Breiman pada tahun 2001 yang merupakan kombinasi dari pohon klasifikasi sedemikian rupa sehingga setiap pohon bergantung pada nilai acak vektor sampel secara mandiri dan dengan distribusi yang sama untuk semua pohon di hutan. *Random Forest* telah banyak digunakan baik untuk klasifikasi dan regresi karena kinerjanya yang unggul dan strukturnya yang sederhana [14]. *Ensemble learning* adalah sebuah komposisi dari beberapa *learner* atau *classifier* yang lemah dengan sebuah *learner* yang memiliki hasil prediksi yang bagus [16].

Ada tiga aspek penting dalam metode *Random Forest:*

1. Melakukan *boostrap sampling* untuk membangun pohon prediksi.
2. Masing-masing pohon keputusan memprediksi dengan prediktor acak.
3. Lalu *Random Forest* melakukan prediksi dengan mengombinasikan hasil dari setiap pohon keputusan dengan cara *majority vote* untuk klasifikasi atau rata-rata untuk regresi [17].

Penggunaan *Random Forest Regression* diharapkan dapat mencapai hasil analisis prediksi *rating* film yang akurat dengan persentase yang tinggi. Juga penggunaan model ini bertujuan untuk mengurangi *well trained* ketika dibentuk dengan *training dataset* untuk mengombinasikan kemudahan *Decision Tree* dengan hasil yang lebih fleksibel.

2.6. Menghitung Pengujian

Tahap pengujian ini dilakukan untuk mendapatkan metrik performa dari model *Random Forest Regression.* Pengujian akan menggunakan teknik *10-fold cross-validation. K-fold cross-validation* adalah salah satu dari jenis pengujian *cross validation* yang berfungsi untuk menilai kinerja proses sebuah metode algoritma dengan membagi sampel data secara acak dan mengelompokkan data tersebut sebanyak nilai *K-fold*. Kemudian salah satu kelompok *K-fold* tersebut akan dijadikan sebagai data uji sedangkan sisa kelompok yang lain akan dijadikan sebagai data latih [18]. Persamaan (1) merupakan contoh penulisan persamaan untuk menguji pola klasifikasi pada penelitian dengan menggunakan teknik *K-fold cross validation* [19].

$$\text{Akurasi} = \frac{jumlah\ klasifikasi\ benar}{jumlah\ data\ uji} \times 100\% \quad (1)$$

Pada penelitian akan melakukan evaluasi menggunakan metrik evaluasi statistik untuk menilai kinerja model. 4 metrik statistik yang digunakan meliputi *coefficient of determination* (R2), *mean squared error* (MSE), *root mean squared error* (RMSE), *mean absolute error* (MAE) [20].

3. HASIL DAN PEMBAHASAN

a. *Dataset*

Pada penelitian ini tidak dilakukan pengumpulan data secara manual atau melakukan tahap proses *web scraping* menggunakan *Application Programming Interface* (API). *Dataset* yang didapatkan dari *website Kaggle.* Pada akun Shakir, J diperoleh *dataset* TMDB dengan *file* berekstensi CSV bernamakan *tmdb_movies* dengan jumlah *rows* sebanyak 11333 dan jumlah atribut sebanyak 17. Tabel 1 menampilkan nama atribut dan tipe data dari *dataset* TMDB.

Table 1. Nama dan Tipe Data Atribut

Nama Atribut	Tipe Data
id	*int64*
imdb_id	*Object*
budget	*int64*
revenue	*int64*
original_title	*Object*
cast	*Object*
homepage	*Object*
director	*Object*
tagline	*Object*
keywords	*Object*
runtime	*float64*
genres	*Object*
production_companies	*Object*
release_date	*Object*
release_year	*float64*
vote_count	*float64*
vote_average	*float64*

b. Hasil *Preprocessing*

Setelah menjalankan proses *data preprocessing* yang meliputi pembersihan data dari *missing values*, *irrelevant*, *inconsistent*, penghapusan *outlier*, pengubahan format *categorical data* menjadi *numerical data* dan mempersiapkan *dataset*. Hasil pada proses *data preprocessing* akan terjabarkan secara rinci dan sistematis pada penjelasan dibawah ini.

1. *Data Cleaning*

Pada tahap ini yang pertama dilakukan ialah mencari *missing values* pada *dataset*. Dikarenakan pada beberapa atribut ditemukan kurang lebih 50% isinya mengandung *missing values*, maka dilakukan pembuangan atribut pada *dataset*. Atribut yang dihapus ialah atribut *imdb_id*, *homepage*, *tagline* dan *keywords*. Pada tabel 2 ditampilkan hasil dari pengecekan *missing values*. Tabel ini menampilkan jumlah *missing values* pada masing-masing atribut.

Tabel 2. *Missing Values* pada Atribut

Nama Atribut	Total *Missing Values*
id	0
imdb_id	278
budget	0
revenue	0
original_title	0
cast	67
homepage	8197
director	26
tagline	3092
keywords	1766
runtime	346
genres	366
production_companies	1317
release_date	616
release_year	342
vote_count	618
vote_average	347

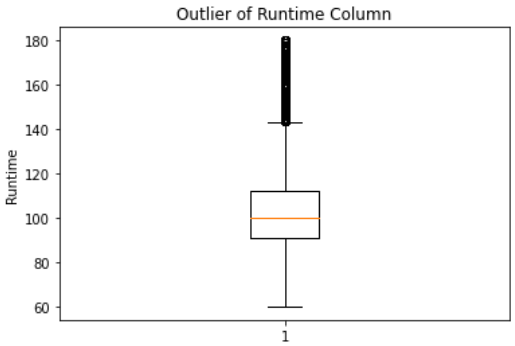
Selain itu penghapusan atribut *release_date* dilakukan karena memiliki kemiripan *values* dengan atribut *release_year*. Juga pada *values* atribut *vote_count* dengan atribut *vote_average* yang akan menjadi *rating* pada penelitian ini. Pada tahap ini atribut berkurang menjadi 9 atribut. Untuk mempermudah pemanggilan atribut pada penelitian, penulis mengubah beberapa nama atribut. Pada tabel 3 diperlihatkan perubahan nama pada atribut *dataset*.

Tabel 3. Perubahan nama atribut

Nama Atribut Lama	Nama Atribut Baru
id	*id*
original_title	*title*
cast	*cast*
director	*director*
runtime	*runtime*
genres	*genres*
production_companies	*pro_com*
release_year	*year*
vote_average	*rating*

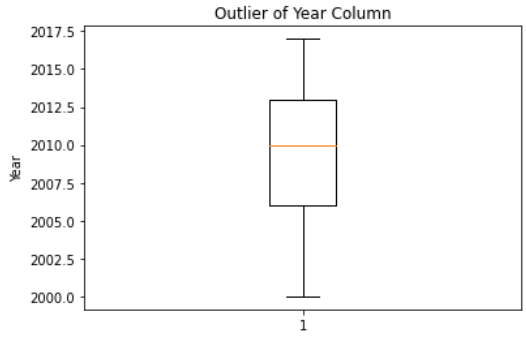
Pada tahap pengecekan duplikasi data ditemukan bahwa pada atribut *title* terdapat 265 *title* yang sama. Dilakukan pengecekan lebih mendalam terhadap *title* yang sama. Ditemukan hasil bahwa secara keseluruhan *title* film hanya memiliki kesamaan atas judul saja. Setelah *dataset* bersih dari *irrelevant*, *inconsistent* dan *missing values*, dilakukan pengubahan format dari atribut. Atribut *cast*, *director* dan *pro_com* dikembangkan menjadi beberapa atribut baru yaitu *cast1*, *cast2*, *cast3*, *director1*, *genre1*, *genre2*, *pro_com1* dan *pro_com2*.

Pada tahap ini dilakukan pengecekan *outlier* pada *dataset*. Pada gambar 2, 3 dan 4 digambarkan hasil pengecekan *outlier* pada atribut *runtime*, *year* dan *rating*. Pada gambar 2 terdapat banyak *outlier* di atribut *runtime*. Untuk film dengan durasi film melebih 140 menit adalah *outlier* pada dataset ini. Sehingga penelitian ini menetapkan akan menganalisis prediksi *rating* film dengan durasi direntang 60 menit sampai 180 menit.



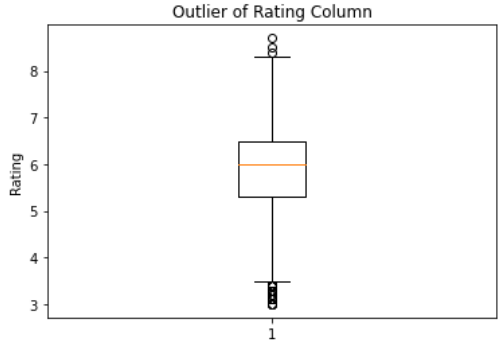
Gambar 2. Visualisasi Hasil *Handling Outlier* pada Atribut *Runtime*

Pada gambar 3 dibawah tidak terdapat *outlier* diluar dari rentang tahun. Namun ditetapkan bahwa rentang tahun penelitian ini dari tahun 2000 sampai tahun 2017.



Gambar 3. Visualisasi Hasil *Handling Outlier* pada Atribut *Year*

Pada gambar 4 digambarkan hasil pengecekan *outlier* pada atribut *rating.* Disimpulkan bahwa *rating* film dibawah 4 dan *rating* film diatas 8 merupaka *outlier* pada *dataset* TMDB.



Gambar 4. Visualisasi Hasil *Handling Outlier* pada Atribut *Rating*

2. *Data Transformation*

Pada tahap ini dilakukan pengubahan *categorical data* menjadi *numerical data.* Beberapa atribut *dataset* mengandung *categorical data* seperti pada atribut *cast1, cast2, cast3, director1, genre1, genre2, pro_com1* dan *pro_com2.* Penelitian ini memutuskan untuk menggunakan teknik *M-estimate encoding.* Pada *M-estimate encoding* akan ditambahkan *smoothing* sehingga untuk *values* yang jarang muncul atau mungkin baru muncul pada *testing data* akan dilakukan *smoothing values* sesuai ketentuan nilai M yang dimasukkan. Sehingga hasil perubahan *categorical data* ke bentuk *numerical data* tidak mengalami *overfitting.* Nilai M pada penelitian ditetapkan sebesar 2. Tabel 3 menjelaskan perubahan dari bentuk *categorical data* menjadi *numerical data* pada *dataset* TMDB. Seperti pada Tabel 3 sebelumnya *cast1* adalah Chris Pratt diubah formatnya menjadi bentuk nominal senilai dengan 6.785518. Ini disimpulkan bahwa *rating* dari Chris Pratt adalah 6.785518.

Tabel 3. Hasil *M-estimate encoding*

Atribut	***Categorical Data***	***Numerical Data***
id	135397	135397
runtime	124	124
year	2015	2015
rating	6.5	6.5
cast1	Chris Pratt	6.785518
cast2	Bryce Dallas Howard	6.042621
cast3	Irrfan Khan	6.402621
director1	Colin Trevorrow	6.502621
genre1	*Action*	5.742755
genre2	*Adventure*	5.839196

Setelah pengubahan *categorical data* menjadi *numerical data*, dilakukan uji korelasi pada atribut-atribut. Pengujian dilakukan dengan menggunakan *Pearson Correlation Coefficient.* Pada pengujian ini semakin mendekati nilai 1 maka korelasi antar atribut semakin baik atribut digunakan untuk penelitian. Berbanding terbalik semakin mendekati -1 maka korelasi antar atribut semakin buruk untuk performa model yang digunakan dalam penelitian ini. Gambar 5 didapatkan hasil korelasi antar atribut, nilai tertinggi pada gambar adalah 0.83 untuk korelasi atribut *rating* dan atribut *director1*. Hal ini disimpulkan bahwa atribut *director1* berperan besar terhadap besarnya *rating* film.

	runtime	year	rating	cast1	cast2	cast3	director1	genre1	genre2	pro_com1	pro_com2
runtime	1.00	-0.05	0.32	0.32	0.28	0.27	0.34	0.16	0.11	0.22	0.25
year	-0.05	1.00	0.04	0.06	0.05	0.03	0.05	0.01	-0.04	-0.04	0.04
rating	0.32	0.04	1.00	0.76	0.80	0.82	0.83	0.39	0.27	0.68	0.62
cast1	0.32	0.06	0.76	1.00	0.67	0.67	0.70	0.34	0.23	0.58	0.51
cast2	0.28	0.05	0.80	0.67	1.00	0.71	0.71	0.36	0.25	0.60	0.52
cast3	0.27	0.03	0.82	0.67	0.71	1.00	0.71	0.37	0.25	0.59	0.52
director1	0.34	0.05	0.83	0.70	0.71	0.71	1.00	0.36	0.24	0.64	0.57
genre1	0.16	0.01	0.39	0.34	0.36	0.37	0.36	1.00	0.18	0.35	0.24
genre2	0.11	-0.04	0.27	0.23	0.25	0.25	0.24	0.18	1.00	0.25	0.16
pro_com1	0.22	-0.04	0.68	0.58	0.60	0.59	0.64	0.35	0.25	1.00	0.49
pro_com2	0.25	0.04	0.62	0.51	0.52	0.52	0.57	0.24	0.16	0.49	1.00

Gambar 5. Visualisasi *Pearson Correlation Coefficient*

c. Pembagian *Dataset*

Penelitian ini membagi keseluruhan *dataset* menjadi 2 bentuk *dataset.* 80% untuk menjadi *training dataset* yang akan jadi bahan pembelajaran pada model *Random Forest Regression.* Dan 20% untuk menjadi *testing dataset* yang digunakan untuk menilai performa dari model yang sudah dibuat. Diperlukan penyesuaian *splitting data* dengan menyatakan parameter *random state* untuk menghindari *training data* dan *testing data* berubah

setiap kali melakukan *running code*. Pada penelitian ini *random state* dinyatakan bernilai 42.

d. Hasil Pengujian

Pada tahap pembuatan model *Random Forest Regression* ditetapkan untuk mencoba menggunakan parameter *n_estimator* sebesar 1000. Dan menetapkan parameter *random_state* sebesar 42. Dalam penelitian ini akan menilai performa model *Random Forest Regression* dengan menggunakan metrik *regression* yaitu *Coefficient of Determination* (R2), *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE) dan 10-*fold cross-validation*.

Tabel 4 ditampilkan hasil dari pengujian dengan metrik statistika dan 10-*fold cross-validation* pada *Random Forest Regression* yang sudah dimodelkan. Pada Tabel 4 untuk hasil R2 dijelaskan bahwa besar koefisien determinasi adalah 86,2%. Dikarenakan nilai R2 mendekati angka 1 atau 100%, maka dapat dikatakan bahwa *independent variables* memberikan pengaruh yang sangat besar pada *dependant variable* pada *dataset* TMDB. Sedangkan untuk MSE dan RMSE mengevaluasi kesalahan prediksi, Tabel 4 menunjukkan rata-rata kesalahan prediksi yang rendah pada model *Random Forest Regression* yaitu hanya sebesar 12% untuk MSE dan kuadrat dari rata-rata kesalahan hanya sebesar 35%. MAE memberikan perbedaan absolut rata-rata antara nilai prediksi dengan nilai aktual pada *testing data*. Pada Tabel 4 nilai metrik MAE bernilai sebesar 22% sehingga terbuktikan rendah nilai MAE semakin baik model *Random Forest Regression* dengan *dataset* TMDB. Dan hasil dari evaluasi performa model dengan *10-fold cross-validation* sebesar 85%.

Tabel 4. Hasil Uji Validasi *Random Forest Regression*

R2	**MSE**	**RMSE**	**MAE**	**10-*fold cross-validation***
0.862	0.12	0.35	0.22	0.85

e. Hasil Uji Pada Berbagai Ukuran *Training Dataset*

Pada penelitian ini dilakukan percobaan membuat model *Random Forest Regression* dalam berbagai ukuran *rows* dari *training dataset.* Pengujian pada percobaan ini dilakukan dengan berbagai bentuk ukuran dengan menggunakan *testing data* yang ukurannya tetap yaitu sebesar 1227 *rows*. Pengujian model *Random Forest Regression* ini hanya akan menggunakan metrik R2. Hasil dari pengujian percobaan ini dijabarkan seperti pada Tabel 5. Dapat disimpulkan selisih pada setiap jumlah *rows training data* tidak terlalu jauh berbeda. Namun semakin banyak jumlah *rows training data* untuk pembelajaran pada model semakin baik performa dari model itu sendiri untuk melakukan prediksi terhadap *rating* film. Pada ukuran *rows* 3000 sampai 4907 nilai R2 tidak mengalami kenaikan. Perlu diteliti lebih lanjut untuk mengetahui penyebab dari samanya nilai performa dari model.

Tabel 5. Hasil Percobaan dengan Ukuran *Training Dataset* yang Beragam

Ukuran *Rows Training Dataset*	**Nilai R2**
200 *rows*	0.834
400 *rows*	0.837
600 *rows*	0.848
800 *rows*	0.850
1000 *rows*	0.852
1500 *rows*	0.857
2000 *rows*	0.861
2500 *rows*	0.861
3000 *rows*	0.862
3500 *rows*	0.862
4000 *rows*	0.862
4907 *rows*	0.862

f. *Hyperparameter Tuning Random Forest Regression*

Random Forest Regression pada *library Scikit Learn* memiliki banyak parameter untuk bisa mengoptimalkan modifikasi dalam pembentukan model prediksi. Parameter yang dimiliki *Random Forest Regression* yang akan dilakukan modifikasi untuk mencari bentuk model yang paling optimal. Pada penelitian ini adalah *max_depth*, *max_features*, *min_samples_leaf*, *min_samples_split*, *n_estimators*.

Dari proses pencarian parameter model *Random Forest Regression* dengan *Grid Search* didapatkan hasil modifikasi terbaik parameter adalah *max_depth* dengan nilai 80, *min_samples_leaf* dengan nilai sebesar 4, *min_samples_split* dengan nilai yang diperoleh sebesar 3 dan *n_estimators* dengan nilai yang diperoleh sebesar 800 pohon.

4. DISKUSI

Dalam penelitian terdahulu yang telah dilakukan sebelumnya. Menurut penelitian yang melakukan perbandingan aplikasi prediksi rating film dengan menggunakan metode *Naive Bayes* dan KNN. Penelitian yang dilakukan dengan menggunakan *dataset* oleh Yueming dari *Northeastern University* disimpulkan bahwa metode

Naive Bayes dengan tingkat keakuratan sebesar 53,99% lebih cocok untuk mengetahui prediksi *rating* film [21].

Pada penelitian yang melakukan prediksi *rating* film *Hollywood* menggunakan model *K-Nearest Neighbor*. Model KNN dapat menghasilkan data yang efektif dan memiliki konsistensi yang kuat. Dari penelitian hasil *accuracy* KNN perlu ditambahkan seleksti fitur *Forward Selection* dan *Backward Elimination*. Dengan seleksi fitur hasil *accuracy* meningkat menjadi 98,92%, *recall* 98,00% dan *precision* 100,00% [22].

Pada penelitian yang mencoba mengoptimalkan analisis sentimen pada ulasan film dengan metode *Naive Bayes* dan *Random Forest* dengan seleksi fitur *Chi-Square*, algoritma *AdaBoost* dan *Voting* memberikan hasil akurasi yang lebih baik. Hasil pengujian pada metode *Naive Bayes* dengan ditambahkan seleksi fitur *Chi-Square* mendapatkan nilai akurasi 79,5%. Untuk pengujian algoritma *AdaBoost* dengan kombinasi model *Random Forest* memiliki hasil pengujian 83,25%. Dibandingkan *AdaBoost* dengan *Naive Bayes* yang hanya memiliki hasil pengujian 81,65%. Untuk hasil algoritma *Voting* 1 dan 2 memiliki nilai 84,6%, *Voting* 2 dengan *Random Forest* dan SVM. Model yang diusulkan dengan *Chi-Square* dan *Voting* 2 dengan *Random Forest* dan SVM memiliki nilai 84,6%. Sehingga kombinasi ini yang terbaik [23].

5. KESIMPULAN

Pada penelitian yang sudah dilaksanakan dengan penggunaan *M-estimate encoding* dalam pengubahan bentuk *categorical data* menjadi bentuk *numerical data* pada atribut *cast1, cast2, cast3, director1, genre1, genre2, pro_com1 dan pro_com2* memberikan solusi pada data-data yang jarang sekali muncul ataupun pada data yang tidak terdaftar pada *training data* sehingga nilai bobot setiap data tidak terlalu tinggi atau bias. Karena dilakukan perhitungan *smooting values* dengan menetapkan nilai M.

Dari hasil proses uji korelasi dengan *Pearson Coefficient Correlation* diperoleh 4 atribut yaitu atribut *director1* dengan perolehan nilai *correlation* tertinggi yaitu sebesar 0.83, disusul dengan atribut *cast3* dengan nilai 0.82. Pada posisi ketiga ada atribut *cast2* dengan nilai 0.8 dan diposisi keempat adalah atribut *cast1* dengan nilai 0.76. Hasil dari pencarian *Grid Search* didapatkan model dengan parameter yang optimal dan lebih akurat yaitu dengan penyesuaian parameter *max_depth* dengan nilai 80, *min_samples_leaf* dengan nilai 4, *min_samples_split* dengan nilai 3, dan parameter *n_estimator* dengan nilai 64 pada 800 pohon.

Pada pengujian dan pengevaluasian performa model *Random Forest Regression* dengan metrik *R2 Score* sebesar 86% untuk jumlah *training data* sebanyak 4907 *rows*. Diperoleh nilai model dengan metrik MSE sebesar 12%. Untuk peroleh nilai dari metrik RMSE adalah sebesar 35%. Dan untuk peroleh metrik *regression* dari MAE adalah sebesar 22%. Dan perolehan dari pengevaluasian performa dengan teknik *10-fold cross-validation* didapatkan rata-rata nilai adalah sebesar 85%.

DAFTAR PUSTAKA

[1] C. Bruneel *et al.*, "Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success," *Nord. J. Media Manag. Issue*, vol. 1, no. 3, pp. 339–359, 2020, doi: 10.5278/njmm.2597-0445.5871.

[2] Y. Pusparisa, "Pertumbuhan Layar dan Bioskop Indonesia Tersendat Pandemi," *30 Maret*, 2021. https://databoks.katadata.co.id/media/statistik/f338238999e2a2c/pertumbuhan-layar-dan-bioskop-indonesia-tersendat-pandemi (accessed Nov. 10, 2022).

[3] S. Sadya, "Daftar Film Indonesia yang Paling Banyak Ditonton pada 2022," *20 Desember*, 2022. https://dataindonesia.id/varia/detail/daftar-film-indonesia-yang-paling-banyak-ditonton-pada-2022 (accessed Nov. 10, 2022).

[4] D. Muhamad Furqon, R. Ahmad Maulana, A. Fauzi, N. Dwi Cahya, M. Nur Sidiq, and T. Kurnia Sandi, "Prediksi Film Pilihan Penonton berdasarkan Genre, Aktor, dan Sutradara Berbasis Data Mining menggunakan Algoritma Eclat (Viewer Movie Predictions based on Genres, Actors, and Directors based on Data Mining Using the Eclat Algorithm)," *Gunung Djati Conf. Ser.*, vol. 3, 2021.

[5] Z. Mhowwala, A. R. Sulthana, and S. D. Shetty, "Movie rating prediction using ensemble learning algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 383–388, 2020, doi: 10.14569/IJACSA.2020.0110849.

[6] R. A. Abarja, "Movie Rating Prediction using Convolutional Neural Network based on Historical Values," *Int. J. Emerg. Trends Eng. Res.*, vol. 8, no. 5, pp. 2156–2164, 2020, doi: 10.30534/ijeter/2020/109852020.

[7] A. V, E. G. Job, N. Sam, and S. M. Sebastian, "Movie success prediction using data mining," *GRD J. Eng.*, vol. 4, no. May, pp. 213–214, 2019.

[8] O. I. Winanda and S. A. Zega, "Prediksi Rating Film Animasi Berdasarkan Elemen," *J. Appl. Multimed. Netw.*, vol. 1, pp. 15–26, 2019.

[9] V. R. Prasetyo, M. Mercifia, A. Averina, L. Sunyoto, and Budiarjo, "Prediksi Rating Film Pada Website IMDB Menggunakan Metode Neural Network," *J. Ilm. Nero*, vol.

7, no. 1, pp. 1–8, 2022.

[10] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart J.*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[11] "Let's talk about TMDB." https://www.themoviedb.org/about.

[12] E. H. S. Atmaja, "Prediksi Kemenangan eSport DOTA 2 Berdasarkan Data Pertandingan," *Avitec*, vol. 2, no. 1, pp. 1–8, 2020, doi: 10.28989/avitec.v2i1.612.

[13] F. Mu'Alim and R. Hiday, "Implementasi Metode Random Forest Untuk Penjurusan Siswa Di Madrasah Aliyah Negeri Sintang," *Jupiter*, vol. 14, no. 1, pp. 116–125, 2022, [Online]. Available: https://www.neliti.com/publications/441871/implementasi-metode-random-forest-untuk-penjurusan-siswa-di-madrasah-aliyah-nege#cite.

[14] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *J. Inform.*, vol. 5, no. 2, pp. 175–185, 2018, doi: 10.31311/ji.v5i2.4158.

[15] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[16] R. Sudiyarno, A. Setyanto, and E. T. Luthfi, "Peningkatan Performa Pendeteksian Anomali Menggunakan Ensemble Learning dan Feature Selection," *Creat. Inf. Technol. J.*, vol. 7, no. 1, p. 1, 2021, doi: 10.24076/citec.2020v7i1.238.

[17] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[18] D. Cahyanti, A. Rahmayani, and S. A. Husniar, "Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 39–43, 2020, doi: 10.33096/ijodas.v1i2.13.

[19] S. Wahyuningsih and D. Retno Utari, "Perbandingan Metode K-Nearest Neighbor, Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit," *Konf. Nas. Sist. Inf. 2018 STMIK Atma Luhur Pangkalpinang*, pp. 8–9, 2018, [Online]. Available: http://jurnal.atmaluhur.ac.id/index.php/knsi2018/article/view/424.

[20] Z. Lyu *et al.*, "Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam," *Materials (Basel).*, vol. 15, no. 4, 2022, doi: 10.3390/ma15041477.

[21] A. R. Yosafat and Y. Kurnia, "Aplikasi Prediksi Rating Film dengan Perbandingan Metode Naïve Bayes dan KNN Berbasis Website Menggunakan Framework Codeigniter," *J. ALGOR*, vol. 1, no. 1, pp. 16–26, 2019, [Online]. Available: https://jurnal.ubd.ac.id/index.php/algor/article/view/221.

[22] A. Bode, "Seleksi Fitur Untuk Prediksi Rating Film Hollywood Menggunakan Model K-Nearest Neighbor," *JUPITER J. Penerapan Ilmu-ilmu Komput.*, vol. 5, no. 1, 2019, [Online]. Available: https://ejournal.borobudur.ac.id/index.php/08/article/view/564.

[23] A. Andreyestha and A. Subekti, "Analisa Sentiment Pada Ulasan Film Dengan Optimasi Ensemble Learning," *J. Inform.*, vol. 7, no. 1, pp. 15–23, 2020, doi: 10.31311/ji.v7i1.6171.