JSAI: Journal Scientific and Applied Informatics
Vol. 5, No. 3, November 2022, hal. 203~208
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085

Prediksi Rating Film Menggunakan Bayesian Regressor dan Gradient Boosting Regressor

[1]**Umniy Salamah**
[1]Universitas Mercu Buana, Indonesia
[1]*umniy.salamah@mercubuana.ac.id*

Article Info

Article history:

Received, 2022-07-28
Revised, 2022-10-25
Accepted, 2022-11-09

Kata Kunci:

Movie Rating
Bayesian Regression
Gradient Boosting Regressor

ABSTRAK

Salah satu fitur yang cukup banyak dikembangkan untuk aplikasi adalah fitur penilaian pengguna. Informasi tentang peringkat pengguna ini dapat digunakan untuk memberikan rekomendasi terbaik tentang hal menarik bagi pengguna lainnya. Sebagai contoh, layanan untuk penjualan film, fitur ini dapat digunakan untuk memberikan rekomendasi yang sesuai dengan peringkat pengguna dan mendorong peningkatan penjualan. Adapun tahapan penelitian adalah Data Preprocessing, Feature Engineering, Modelling dan Evaluation. Penelitian ini menggunakan metode yaitu Bayesian Regressor dan Gradient Boosting Regressor untuk memprediksi movie rating. Penelitian ini menggunakan TMDB 5000 Movie Dataset yang terdiri dari kurang lebih 4800 data. Sebagai hasilnya, Gradient Boosting Regressor memiliki hasil yang lebih baik dibandingkan Bayesian Ridge Regressor. Gradient Boosting Regressor memiliki nilai R^2 score sebesar 0.843.

Keywords:

Movie Rating
Bayesian Regression
Gradient Boosting Regressor

ABSTRACT

One of the features that have been widely developed for applications is the user rating feature. This information about user ratings can be used to provide the best recommendations about things of interest to other users. For example, a service for selling movies, this feature can be used to provide recommendations according to user ratings and drive increased sales. The research stages are Data Preprocessing, Feature Engineering, Modeling, and Evaluation. This study uses the Bayesian Regressor and Gradient Boosting Regressor methods to predict movie ratings. This study uses the TMDB 5000 Movie Dataset, which consists of approximately 4800 data. As a result, the Gradient Boosting Regressor has better results than the Bayesian Ridge Regressor. Gradient Boosting Regressor has an R ^ 2 score of 0.843.

Penulis Korespondensi:

Umniy Salamah
Fakultas Ilmu Komputer
Universitas Mercu Buana, Indonesia
Email: umniy.salamah@mercubuana.ac.id

1. PENDAHULUAN

Pada era teknologi informasi dan komunikasi saat ini, pengembang sistem atau aplikasi telah memikirkan cara menyampaikan layanan dan informasi yang dapat memberikan pengalaman terbaik bagi pengguna. Salah satu fitur yang cukup banyak dikembangkan untuk aplikasi adalah fitur penilaian pengguna. Penelitian ini dapat digunakan untuk menentukan peringkat suatu hal, barang atau produk [1]–[6]. Penilaian pengguna ini dapat dikembangkan dengan cara memberikan rekomendasi peringkat secara otomatis. Perhitungan peringkat rekomendasi berdasarkan karakteristik yang pada hal, barang atau produk yang dinilai [2], [3], [7]–[12].

Informasi tentang peringkat pengguna ini dapat digunakan untuk memberikan rekomendasi terbaik tentang hal menarik bagi pengguna lainnya [13]. Sebagai contoh, layanan untuk penjualan film, fitur ini dapat digunakan untuk memberikan rekomendasi yang sesuai dengan peringkat pengguna dan mendorong

peningkatan penjualan. Fitur ini sangat berguna untuk menarik perhatian pengguna dan meningkatkan kualitas layanan dari aplikasi tersebut [2], [14], [15].

Data interpretasi yang tepat dari data dari pengguna terhadap objek yang dinilai akan meningkatkan prediksi peringkat penilaian menjadi lebih baik [16]. Selain itu, perlu pemilihan metode yang sesuai terhadap karakteristik data yang diproses perlu dipertimbangkan untuk membangun fitur prediksi yang akurat. Metode yang digunakan dapat bervariasi tergantung dengan jumlah data dan karakteristik dari objek yang dinilai. [2], [17]–[19].

Penelitian mengenai prediksi penilaian ini telah dilakukan oleh beberapa peneliti, antara lain: [20]–[24]. Penelitian oleh Jeon et al. (2009) menggunakan algoritma gradient boosting machine (XGBoost) untuk melakukan rating prediction pada objek movie. Penelitian ini menggunakan MovieLens20M dataset. Penelitian ini mendapatkan nilai MAE (Mean Absolute Error) sebesar 0.63399 dan nilai RMSE (Root-mean Squared Error) sebesar 0.82913 dalam melakukan movie rating prediction [22].

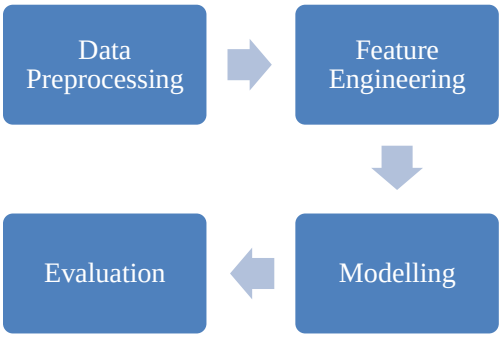
Penelitian oleh Purnomo & Endah (2019) menggunakan *dissymmetrical percentage collaborative filtering algorithm* (DSPCFA) dan collaborative filtering algorithm (CFA) untuk memprediksi movie ratings. Dari hasil penelitian, metode DSPCFA memperoleh nilai error yang lebih rendah dibandingkan metode CFA (perbedaaannya 5% untuk nilai RMSE (Root-mean Squared Error) dan 7% untuk nilai MAE (Mean Absolute Error) [23].

Penelitian oleh YuMin et al. (2018) menggunakan metode KNN, decision tree, SVM dan NBC untuk memprediksi movie rating. Sebagai hasilnya, penelitian ini memperoleh akurasi sebesar 89.8% untuk metode KNN, 48.0% untuk metode decision tree, 50.5% untuk SVM and 37.2% untuk NBC [24].

Berdasarkan latar belakang penelitian diatas, studi ini bertujuan untuk menggunakan metode lain yaitu Bayesian Regressor dan Gradient Boosting Regressor dalam memprediksi movie rating. Penelitian ini menggunakan TMDB 5000 Movie Dataset yang terdiri dari kurang lebih 4800 data.

2. METODE PENELITIAN

Penelitian ini merupakan penelitian benchmark dengan menggunakan dataset public. Dataset yang digunakan adalah TMDB 5000 Movie Dataset. Adapun tahapan penelitian adalah Data Preprocessing, Feature Engineering, Modelling dan Evaluation seperti yang ada pada Gambar berikut ini.



Gambar 1 Tahapan Penelitian

1. Preprocessing data

Pada pra-pengolahan data tahapan yang dilakukan dengan cara memformat fitur, menghapus data yang memiliki fitur kosong, dan menghapus fitur yang tidak diperlukan. Pada dataset, fitur dalam format JSON sehingga akan dikonversi ke dalam format list menggunakan dataframe pada python. Kemudian, jika film tidak memiliki informasi untuk ‘cast’, ‘crew’, dan ‘production_companies’ maka film tidak masuk kedalam kriteria data valid, sehingga akan dihapus dari dataset. Selanjutnya fitur yang tidak diperlukan juga akan dihapus untuk memudahkan dan mempercepat proses komputasi.

2. Feature Engineering

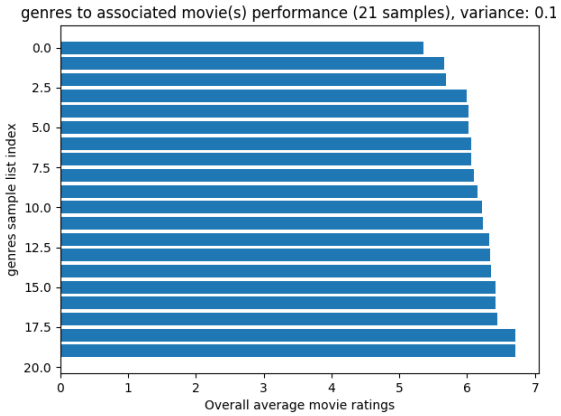
Feature engierring dilakukan dengan mengurutkan fitur dari yang paling tinggi asosiasinya dengan ‘rating’, sampai yang paling rendah asosianya dengan ‘rating’. Dengan mengurutkan fitur, kita dapat melihat fitur yang paling tinggi korelasinya dengan rating film dan dapat membuang fitur yang korelasinya paling rendah dengan rating film. Kemudian, fitur nominal diubah ke bentuk numerik menggunakan teknik pembobotan. Selajutnya, tahap ini melakukan normalisasi fitur menggunakan MinMaxScaler dengan rentang 0-1.

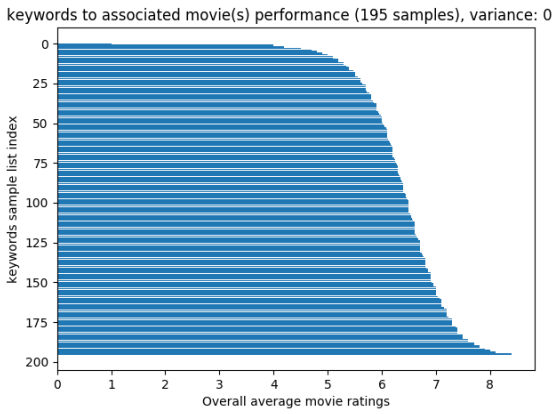
3. Pembangunan Model
Model yang akan diuji coba pada eksperimen ini adalah dua model regresi yaitu Bayesian Ridge Regressor dan Gradient Boosting Regressor. Alasan pemilihan model Bayesian Ridge Regression adalah karena ideal untuk menangani data yang memiliki banyak *outlier* (pencilan). Sedangkan Gradient Boosting merupakan salah satu teknik paling *powerful* untuk membangun model prediksi. Kedua model ini telah disediakan oleh Scikit Learn.

Dataset dibagi menjadi data training dan data testing. Pada tahap ini data akan dibagi menjadi 70% untuk training, 30% untuk testing untuk validasi. Tahap selanjutnya adalah pelatihan model. Pada tahap ini akan dilatih model dalam memetakan fitur dengan target pada data pelatihan. Selanjutnya pada tahap ini akan dilakukan testing pada model yang telah dihasilkan dengan data uji. Testing dilakukan dengan memprediksi target (dalam hal ini 'rating') menggunakan data fitur test.
4. Evaluasi
Evaluasi dilakukan dengan menghitung nilai R^2 score. R^2 score ideal untuk model regressi karena dapat mengukur seberapa baik variasi 'rating' dapat dijelaskan oleh fitur.

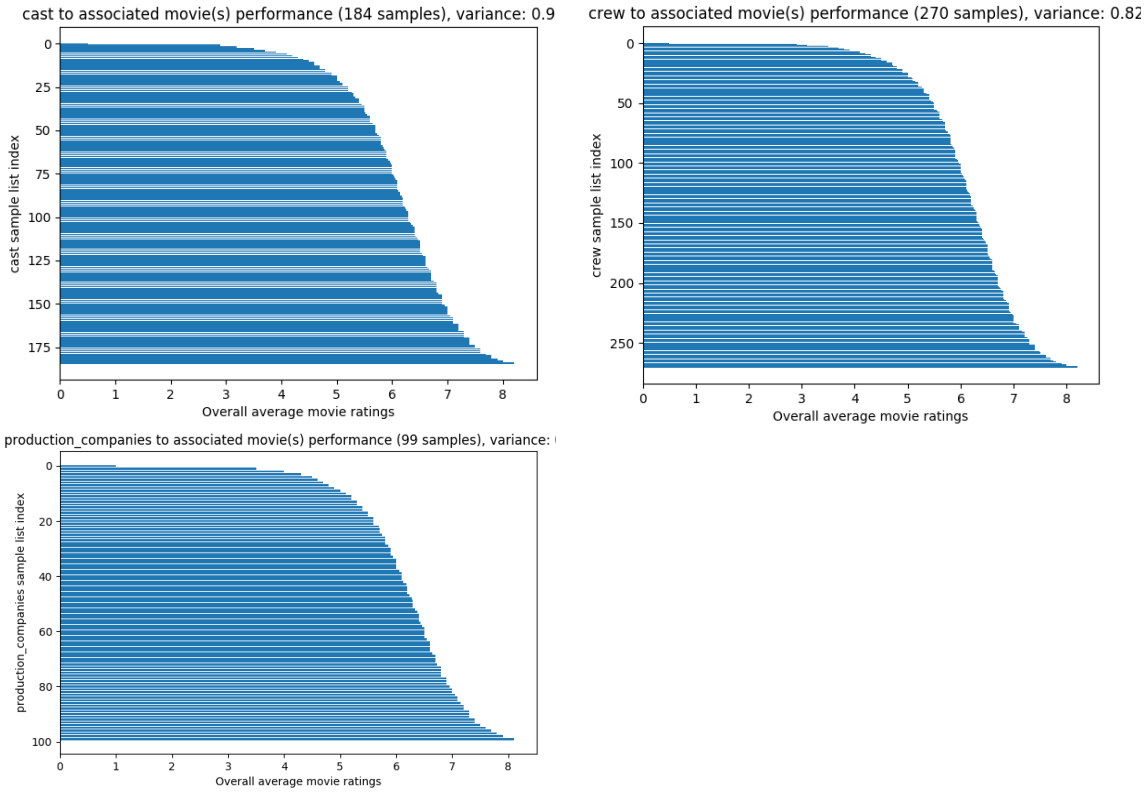
3. HASIL DAN ANALISIS

Salah satu fitur yang cukup banyak dikembangkan untuk aplikasi adalah fitur penilaian pengguna. Informasi tentang peringkat pengguna ini dapat digunakan untuk memberikan rekomendasi terbaik tentang hal menarik bagi pengguna lainnya. Sebagai contoh, layanan untuk penjualan film, fitur ini dapat digunakan untuk memberikan rekomendasi yang sesuai dengan peringkat pengguna dan mendorong peningkatan penjualan. Adapun tahapan penelitian adalah Data Preprocessing, Feature Engineering, Modelling dan Evaluation. Penelitian ini menggunakan metode yaitu Bayesian Regressor dan Gradient Boosting Regressor untuk memprediksi movie rating. Penelitian ini menggunakan TMDB 5000 Movie Dataset yang terdiri dari kurang lebih 4800 data.
Pada tahap preprocessing, fitur yang digunakan pada eksperimen ini adalah: 'genres', 'keywords', 'cast', 'crew', and 'production_companies'. Sedangkan fitur seperti 'id' dan 'original_title' tidak dibutuhkan sehingga dihapus. Pada Gambar 2, memperlihatkan hasil perhitungan asosiasi fitur 'genres', 'keywords', 'cast', 'crew', and 'production_companies' dengan 'rating'. Berdasarkan variance yang dihasilkan fitur 'genre' memiliki variance yang paling rendah, dapat dikatakan bahwa fitur 'genre' tidak berkorelasi dengan 'rating' sehingga tidak akan berguna untuk memprediksi 'rating'. Sedangkan fitur 'cast', 'crew', 'production_companies', dan fitur 'keywords' memiliki variance yang tinggi dan akan berguna untuk memprediksi 'rating'.





JSAI: Journal Scientific and Applied Informatics
Vol. 5, No. 3, November 2022, hal. 203~208
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 5
DOI: 10.36085



Gambar 2 Association of features with Rating

Pada tahap ini dikarenakan fitur ‘genre’ tidak dianggap memiliki korelasi dengan ‘rating’ maka fitur ‘genre’ dihapus dari fitur. Sehingga fitur yang digunakan adalah: ‘cast’, ‘crew’, ‘production companies’, dan ‘keywords’. Berikut hasil perbandingan R^2 score untuk Bayesian Ridge Regressor dan Gradient Boosting Regressor.

Tabel 2 Hasil Penelitian

Model	R^2 score
Bayesian Ridge Regressor	0.801
Gradient Boosting Regressor	0.843

Hasil evaluasi menunjukkan Gradient Boosting Regressor memiliki hasil yang lebih baik dibandingkan Bayesian Ridge Regressor. Gradient Boosting Regressor memiliki nilai cukup baik yaitu nilai R^2 score sebesar 0.843, hal ini menunjukkan bahwa nilai ‘rating’ dapat dipengaruhi oleh fitur (‘cast’, ‘crew’, ‘production companies’, dan ‘keywords’).

4. KESIMPULAN

Salah satu fitur yang cukup banyak dikembangkan untuk aplikasi adalah fitur penilaian pengguna. Informasi tentang peringkat pengguna ini dapat digunakan untuk memberikan rekomendasi terbaik tentang hal menarik bagi pengguna lainnya. Sebagai contoh, layanan untuk penjualan film, fitur ini dapat digunakan untuk memberikan rekomendasi yang sesuai dengan peringkat pengguna dan mendorong peningkatan penjualan. Adapun tahapan penelitian adalah Data Preprocessing, Feature Engineering, Modelling dan Evaluation. Penelitian ini menggunakan metode yaitu Bayesian Regressor dan Gradient Boosting Regressor untuk memprediksi movie rating. Penelitian ini menggunakan TMDB 5000 Movie Dataset yang terdiri dari kurang lebih 4800 data. Sebagai hasilnya, Gradient Boosting Regressor memiliki hasil yang lebih baik dibandingkan Bayesian Ridge Regressor. Gradient Boosting Regressor memiliki nilai R^2 score sebesar 0.843.

UCAPAN TERIMA KASIH

Terima kasih kepada Pusat Penelitian (Biro Penelitian, Pengabdian Masyarakat & Publikasi UMB) yang telah mendanai penelitian ini dengan kontrak penelitian 02-5/891/B-SPK/III/2022

REFERENSI

[1] S. Basu, "Movie rating prediction system based on opinion mining and artificial neural networks," in *International Conference on Advanced Computing Networking and Informatics*, 2019, pp. 41–47.

[2] M. Marović, M. Mihoković, M. Mikša, S. Pribil, and A. Tus, "Automatic movie ratings prediction using machine learning," in *2011 Proceedings of the 34th International Convention MIPRO*, 2011, pp. 1640–1645.

[3] I. Nurhaida, V. Ayumi, D. Fitrianah, R. A. M. Zen, H. Noprisson, and H. Wei, "Implementation of deep neural networks (DNN) with batch normalization for batik pattern recognition," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 2, pp. 2045–2053, 2020.

[4] H. Noprisson, N. Husin, N. Zulkarnaim, P. Rahayu, A. Ramadhan, and D. I. Sensuse, "Antecedent Factors of Consumer Attitudes toward SMS , E-mail and Social Media for Advertising," in *ICACSIS 2016*, 2016.

[5] N. Ani, H. Noprisson, and N. M. Ali, "Measuring usability and purchase intention for online travel booking: A case study," *Int. Rev. Appl. Sci. Eng.*, vol. 10, no. 2, pp. 165–171, 2019.

[6] A. Ratnasari, D. Fitrianah, and W. H. Haji, "BPTrends Redesign Methodology (BPRM) for the Development Disaster Management Prevention Information System," in *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference*, 2020, pp. 113–117.

[7] Y. Devianto and S. Dwiasnati, "Application electronic marketing to help marketing leading products village," *GSC Adv. Eng. Technol.*, vol. 1, no. 1, pp. 65–74, 2021.

[8] E. M. Khan, M. S. H. Mukta, M. E. Ali, and J. Mahmud, "Predicting Users' Movie Preference and Rating Behavior from Personality and Values," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 3, pp. 1–25, 2020.

[9] H. Noprisson, E. Ermatita, A. Abdiansah, V. Ayumi, M. Purba, and H. Setiawan, "Fine-Tuning Transfer Learning Model in Woven Fabric Pattern Classification," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 06, p. 1885, 2022.

[10] H. Noprisson, E. Ermatita, A. Abdiansah, V. Ayumi, M. Purba, and M. Utami, "Hand-Woven Fabric Motif Recognition Methods: A Systematic Literature Review," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS*, 2021, pp. 90–95.

[11] V. Ayumi, E. Ermatita, A. Abdiansah, H. Noprisson, M. Purba, and M. Utami, "A Study on Medicinal Plant Leaf Recognition Using Artificial Intelligence," in *2021 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS*, 2021, pp. 40–45.

[12] I. Nurhaida *et al.*, "Implementation of Deep Learning Predictor (LSTM) Algorithm for Human Mobility Prediction," *Int. J. Interact. Mob. Technol.*, vol. 14, no. 18, p. 132, Nov. 2020.

[13] W. R. Bristi, Z. Zaman, and N. Sultana, "Predicting imdb rating of movies by machine learning techniques," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019, pp. 1–5.

[14] S. Dwiasnati and Y. Devianto, "Utilization of Prediction Data for Prospective Decision Customers Insurance Using the Classification Method of C. 45 and Naive Bayes Algorithms," in *Journal of Physics: Conference Series*, 2019, vol. 1179, no. 1, p. 12023.

[15] K. Pradeep, C. R. TintuRosmin, S. S. Durom, and G. S. Anisha, "Decision Tree Algorithms for Accurate Prediction of Movie Rating," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 853–858.

[16] R. Shrivastava and H. Singh, "K-means clustering based solution of sparsity problem in rating based movie recommendation system," *Int. J. Eng. Manag. Res.*, vol. 7, no. 2, pp. 309–314, 2017.

[17] X. Ning, L. Yac, X. Wang, B. Benatallah, M. Dong, and S. Zhang, "Rating prediction via generative convolutional neural networks based regression," *Pattern Recognit. Lett.*, vol. 132, pp. 12–20, 2020.

[18] B. L. Devi, V. V. Bai, S. Ramasubbareddy, and K. Govinda, "Sentiment analysis on movie reviews," in *Emerging Research in Data Engineering Systems and Computer Communications*, Springer, 2020, pp. 321–328.

[19] M. A. Shahjalal, Z. Ahmad, M. S. Arefin, and M. R. T. Hossain, "A user rating based collaborative filtering approach to predict movie preferences," in *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, 2017, pp. 1–5.

[20] N. Choudhry, J. Xie, and X. Xia, "Big Data Analytics of Movie Rating Predictive System," in *Journal of Physics: Conference Series*, 2020, vol. 1575, no. 1, p. 12063.

[21] S. Gogineni and A. Pimpalshende, "Predicting IMDB Movie Rating Using Deep Learning," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1139–1144.

[22] T. Jeon, J. Cho, S. Lee, G. Baek, and S. Kim, "A movie rating prediction system of user propensity analysis based on collaborative filtering and fuzzy system," in *2009 IEEE International Conference on Fuzzy Systems*, 2009, pp. 507–511.

[23] J. E. Purnomo and S. N. Endah, "Rating prediction on movie recommendation system: collaborative filtering algorithm (CFA) vs. dissymetrical percentage collaborative filtering algorithm (DSPCFA)," in *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)*, 2019, pp. 1–6.

[24] S. YuMin, Z. Yuan, and Y. JinYao, "Neural Network Based Movie Rating Prediction," in *Proceedings of the 2018 International Conference on Big Data and Computing*, 2018, pp. 33–37.