Media Journal of General Computer Science, 2025, Vol 2, Issue 1,12-17
Published Online January 2025 in MASE (http://www.mase.or.id)
DOI: 10.62205/mjgcs.v2i1.35

Optimizing Amazon Reviews Using Principal Component Analysis, Feature Selection on Random Forest Classifier

M Nabil Fadhlurrahman
Department of Informatics, Dinamika Bangsa University, Jambi, Indonesia
E-mail: mnabilf81@gmail.com

Mutiara Yudina Fitrah
Department of Informatics, Dinamika Bangsa University, Jambi, Indonesia
E-mail: mutiarayudhinafitrah@gmail.com
*Correspondent Author

Received: 11 August, 2024; Accepted: 3 November, 2024; Published: 30 January, 2025

Abstract: Dataset optimization is an important step in machine learning to improve model performance. This review discusses the use of Random Forest, Principal Component Analysis (PCA), and Feature Selection algorithms to optimize datasets. Based on this review, the combination of Random Forest, PCA, and Feature Selection is proven to be effective in improving machine learning model performance. This combination can help reduce overfitting, improve prediction accuracy, and speed up the model training process. In our experiments with the Amazon Reviews dataset, this optimized approach achieved an impressive accuracy of 91%, demonstrating a significant improvement over baseline models.

Keywords: *Machine Learning, Random Forest Classifier, PCA, Feature Selection*.

I. Introduction

Menganalisis Dataset Review Amazon dengan Random Forest, PCA, dan FS menghadirkan beberapa tantangan, seperti ketidakkonsistenan ulasan, bias, dimensi data yang tinggi, interpretasi model yang kompleks, pemilihan model dan parameter yang tepat, dan efisiensi komputasi. Mengatasi tantangan ini melalui pembersihan data, deteksi spam, optimasi hyperparameter, sampling, dan pemilihan teknik yang tepat dapat membantu mendapatkan informasi berharga dari dataset tersebut. Dengan ini peneliti ingin mencoba untuk mengoptimasi dataset Amazon Review menggunakan Principal Component Analysis, Feature Selection dengan Random Forest Classifier.

Optimasi dalam machine learning merupakan hal yang sangat penting dalam penggunaan Machine Learning. Optimasi menggunakan PCA dan FS untuk meningkatkan performa dan efektivitas. PCA membantu mengatasi overfitting dan meningkatkan akurasi model, sedangkan FS memilih subset fitur yang relevan dan informatif untuk mengurangi noise dan meningkatkan akurasi. Penerapan metode ini menghasilkan peningkatan akurasi, pengurangan false positive, peningkatan efisiensi, dan pengembangan model yang lebih adaptif. Optimasi dengan PCA dan FS menjadi kunci untuk menyimpulkan hasil dari dataset tersebut secara efektif. Optimasi membantu meminimalkan kesalahan prediksi model, menghasilkan prediksi yang lebih akurat dan andal pada data baru. [1]. Meningkatkan kemampuan model untuk beradaptasi dan menghasilkan prediksi akurat pada data baru, mencegah overfitting dan memastikan model yang robust. [2]. Membantu meningkatkan pemahaman tentang bagaimana model membuat prediksi, menghasilkan solusi yang lebih mudah dipahami dan diinterpretasikan.[3]

Machine learning merupakan salah satu cabang ilmu Kecerdasan Buatan (Artificial Intelligence) yang berkembang sangat cepat dan telah menyebabkan masalah klasifikasi, regresi, klastering, dan anomaly detection pada berbagai bidang dapat diatasi lebih efisien.[4]**.** Luasnya potensi aplikasi machine learning telah menginspirasi banyak peneliti untuk terus mengembangkan model dan teknologi machine learning yang menghasilkan sejumlah besar publikasi ataupun prototipe produk teknologi cerdas. ML juga berperan penting dalam mengidentifikasi hubungan baru antara variabel penelitian. Algoritma ML canggih dapat mendeteksi korelasi dan pola yang kompleks dalam data, mendorong pemahaman yang lebih baik tentang fenomena yang sedang diteliti dan memicu penemuan ilmiah baru [5]. ML menawarkan potensi untuk mengotomatisasi tugas yang berulang dan memakan waktu, seperti pengumpulan data, analisis data, dan pembuatan laporan. Hal ini membebaskan waktu peneliti untuk fokus pada aspek kreatif dan bernilai tinggi dari penelitian mereka, meningkatkan produktivitas dan efisiensi penelitian secara keseluruhan. [6].

Penelitian sebelumnya [7] **t**elah mengusulkan penelitian untuk memahami opini dan sentimen pelanggan terhadap produk atau layanan yang ditawarkan, tantangan pada penelitian ini dalam mengklasifikasikan review dengan tepat terutama untuk review yang ambigu atau sarkasme. Penelitian ini menggunakan Algoritma Random Forest dipilih karena memiliki performa Akurasi dan Recall diantara 4 algoritma lain. Lalu penelitian selanjutnya [8] Juga telah mengusulkan untuk menggunakan Random Forest pada Amazon Review, tantangan pada penelitian ini terdapat pada prediksi rating produk. Terdapat banyak faktor yang mempengaruhi rating seperti jumlah review, kualitas produk, layanan pelanggan, harga dan pengalaman pribadi pembeli. Penelitian ini juga menggunakan Algoritma Random Forest karena performa yang lebih baik dari algoritma lainnya.

Penggunaan Random Forest dalam penelitian [9] menunjukkan potensi besar algoritma ini dalam memahami sentimen pelanggan berdasarkan ulasan produk di Amazon. Dengan kemampuannya mengolah data kompleks dan mengidentifikasi pola yang rumit, Random Forest dapat memberikan hasil klasifikasi yang akurat. Informasi berharga yang didapat dari analisis ini dapat membantu bisnis mengidentifikasi kekuatan dan kelemahan produk, membandingkan kinerja produk yang berbeda, mengidentifikasi tren pasar, serta meningkatkan kepuasan pelanggan. Keunggulan Random Forest seperti kemampuan menangani data tidak seimbang, robust terhadap overfitting, dan fleksibilitas dalam menangani berbagai jenis fitur membuatnya menjadi pilihan yang sangat baik untuk analisis sentimen teks. Hasil dari analisis sentimen menggunakan Random Forest dapat memberikan wawasan yang mendalam bagi perusahaan untuk mengambil keputusan bisnis yang lebih baik.

2. Research Method

2.1. Experiment Setup

Penelitian ini bertujuan untuk mengidentifikasi fitur-fitur apa saja yang mempengaruhi review dan memetakan point-point penting agar dapat menarik kesimpulan dari data dengan efektif. Dalam mencapai tujuan tersebut, peneliti menggunakan algoritma Random Forest, Principal Component Analysis dan Feature Selection sebagai alat menghitung data. Berikut beberapa tahap penelitian yang perlu dilakukan akan diuraikan secara rinci dalam alur penelitian yang ditampilkan dalam Fig 1.

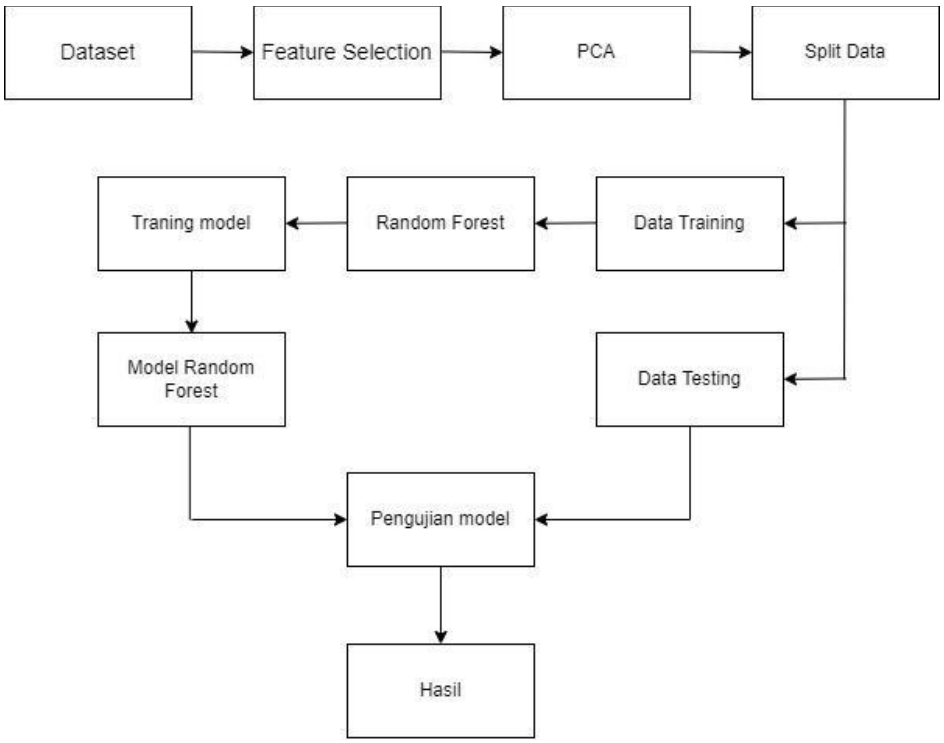

Fig.1. Experiment Setup.

Fig 1 merupakan diagram eksperimen yang menunjukkan tahapan-tahapan yang akan dilakukan dalam penelitian ini dengan melibatkan tiga fase tahapan terpisah, yaitu :

- Mengatur dataset dan mengekstrak fitur-fiturnya menggunakan metode PCA, lalu membagi dataset tersebut menjadi dua bagian: data training dan data testing.
- Pada tahap ini, model akan dilatih menggunakan algoritma Random Forest untuk memperoleh hasil yang akan digunakan dalam pengujian model.
- Pada tahap akhir, model yang telah dilatih akan diuji, dan tingkat keberhasilan peningkatan dalam akurasi,

presisi, serta recall akan dihitung.

2.2. Dataset

Dataset Amazon Review adalah kumpulan besar review pelanggan dan informasi produk dari Amazon. Ini adalah sumber data yang kaya bagi para peneliti dan ilmuwan data yang tertarik dalam pemrosesan bahasa alami, analisis sentimen, sistem rekomendasi, dan berbagai tugas pembelajaran mesin lainnya. Dengan memahami kekuatan dan tantangan dari Dataset Amazon Review, para peneliti dan ilmuwan data dapat memanfaatkan potensinya untuk memperoleh wawasan berharga dan mengembangkan aplikasi inovatif [10].

2.3 Proposed Method

Feature selection sangat penting dalam memilih fitur paling relevan pada dataset Amazon Review. Proses ini mengidentifikasi subset fitur yang paling berkontribusi pada prediksi atau klasifikasi model machine learning. Dengan memilih fitur penting, feature selection menghilangkan fitur tidak relevan atau menambah noise, meningkatkan akurasi prediksi, dan mencegah overfitting. Mengurangi jumlah fitur juga meningkatkan efisiensi komputasi, mempercepat waktu pemrosesan, dan mengurangi penggunaan memori. Ini penting untuk dataset besar seperti Amazon Review. Selain itu, model dengan lebih sedikit fitur lebih mudah diinterpretasi, memberikan wawasan berharga untuk analisis bisnis dan pengambilan keputusan.

Dalam penelitian ini, peneliti menerapkan metode Principal Component Analysis (PCA) untuk meningkatkan akurasi serta mengurangi jumlah dimensi, membantu membuat model lebih efisien dan mengurangi risiko overfitting. PCA mengubah fitur yang saling berkorelasi menjadi set fitur baru yang tidak berkorelasi, disebut principal components. Ini mengurangi jumlah fitur sambil mempertahankan informasi penting dari dataset asli, membuat model lebih sederhana dan mudah diinterpretasi. PCA juga mengidentifikasi dan menggabungkan fitur yang berkorelasi tinggi, mengurangi redundansi dalam data dan meningkatkan kinerja model. PCA dapat mereduksi data ke dua atau tiga dimensi, yang kemudian bisa divisualisasikan untuk analisis lebih lanjut.

Selain itu, peneliti menggunakan Random Forest Classifier, sebuah algoritma ensemble learning yang menggabungkan beberapa pohon keputusan selama pelatihan. Output dari Random Forest adalah modus dari kelas (untuk klasifikasi) atau rata-rata prediksi (untuk regresi) dari masing-masing pohon. Random Forest Classifier sangat efektif dalam menangani dataset besar dan kompleks seperti Amazon Review karena mampu mengelola sejumlah besar fitur dan menangani missing values dengan baik. Algoritma ini juga membantu dalam mengestimasi pentingnya fitur, yang dapat digunakan bersama teknik feature selection dan PCA untuk meningkatkan akurasi model secara keseluruhan.

2.4. Environment Setup

Eksperimen dalam penelitian ini dilaksanakan pada sebuah laptop dengan spesifikasi tertentu, yaitu menggunakan Windows 11, dengan Prosesor Intel Core i5-10300H dengan kecepatan sekitar 2.50GHz serta Ram berkapasistas 16 GB, Beserta alat-alat yang digunakan untuk penelitian meliputi Google Collab, Word, Excel, dan Figma.

3. Result and Discussion

Bagian ini menyajikan hasil eksperimen yang telah dilakukan, termasuk laporan hasil dari langkah-langkah reduksi fitur dan pengujian kinerja algoritma Random Forest. Dalam pembahasan ini, akan dievaluasi efektivitas proses reduksi fitur serta dianalisis kinerja algoritma Random Forest.

3.1 Hasil Pemilihan Fitur

Dalam proses seleksi fitur ini, algoritma Random Forest akan digunakan untuk menangani data dengan dimensi tinggi serta mengurangi risiko overfitting. Selain itu, metode Principal Component Analysis (PCA) akan diterapkan untuk mengurangi jumlah fitur. Tujuan dari pendekatan ini adalah untuk mengurangi beban komputasi dan meningkatkan akurasi serta presisi tanpa kehilangan karakteristik data. Tabel 1 menunjukkan hasil yang dilakukan oleh fitur seleksi dalam menghasilkan tiga fitur penting dalam dataset amazon reviews.

Tabel 1. Hasil Pemilihan Fitur

Fitur	Skor Fitur
summary	0.105430
helpful	0.041077
total_vote	0.031156

reviewText	0.025885
helpful_yes	0.008666
asin	0.006920
unixReviewTime	0.005655
day_diff	0.001824
reviewTime	0.000077

3.2 Penggunaan Random Forest Classifier sebelum menggunakan PCA dan FS

Dalam pengujian awal untuk membuktikan peningkatan akurasi, presisi, dan recall peneliti mencoba menggunakan model Random Forest Classifier dahulu tanpa menggunakan PCA dan Feature Selection. Dilakukan 10 kali pengujian terhadap model ini untuk memastikan konsistensi hasil. Hasil pengujian ini disajikan dalam tabel 2 sebagai berikut :

Tabel 2. Hasil Pengujian Model Random Forest sebelum menggunakan FS dan PCA

Pengujian	Akurasi	Presisi	Recall
1	78%	66%	78%
2	78%	66%	78%
3	79%	67%	79%
4	79%	67%	79%
5	78%	66%	78%
6	78%	66%	78%
7	78%	66%	78%
8	79%	67%	79%
9	79%	67%	79%
10	79%	67%	79%

Berdasarkan hasil tabel pengujian model Random Forest yang dilakukan sebanyak 10 kali sebelum menggunakan Feature Selection dan Principal Component Analysis, dapat disimpulkan bahwa model ini telah menunjukkan hasil akurasi, recall, dan presisi yang cukup baik dengan konsistensi hasil yang stabil. Peneliti berkeinginan untuk meningkatkan akurasi, presisi, dan recall pada dataset ini. Rata-rata akurasi yang di dapatkan dari 10 kali pengujian adalah 78,5%.

3.3 Hasil Optimasi Random Forest menggunakan PCA dan FS

Dalam proses kali ini peneliti telah menerapkan metode Feature Selection dan PCA ke dalam model klasifikasi Random Forest. Sama hal nya dalam pengujian awal, dalam pengujian ini juga dilakukan 10 kali pengujian terhadap model untuk memastikan hasil yang konsisten. Hasil pengujian ini akan disajikan dalam tabel 3 melalui hasil classification report sebagai berikut :

Tabel 3. Hasil Pengujian Model Random Forest setelah menggunakan FS dan PCA

Pengujian	Akurasi	Presisi	Recall
1	91%	86%	91%
2	91%	86%	91%
3	92%	87%	92%
4	91%	86%	91%
5	92%	87%	92%
6	92%	87%	92%
7	92%	88%	92%
8	91%	87%	91%
9	92%	87%	92%
10	91%	86%	91%

Berdasarkan hasil tabel pengujian model Random Forest yang dilakukan sebanyak 10 kali, dapat disimpulkan bahwa model ini menunjukkan hasil yang konsisten dan stabil, dengan akurasi tinggi serta performa yang baik dalam presisi dan recall. Hal ini menunjukkan bahwa metode Feature Selection dan Principal Component Analysis (PCA)

mampu membantu meningkatkan akurasi, presisi, dan recall model Random Forest. Dengan rata-rata akurasi sebesar 91,7% dari 10 kali pengujian.

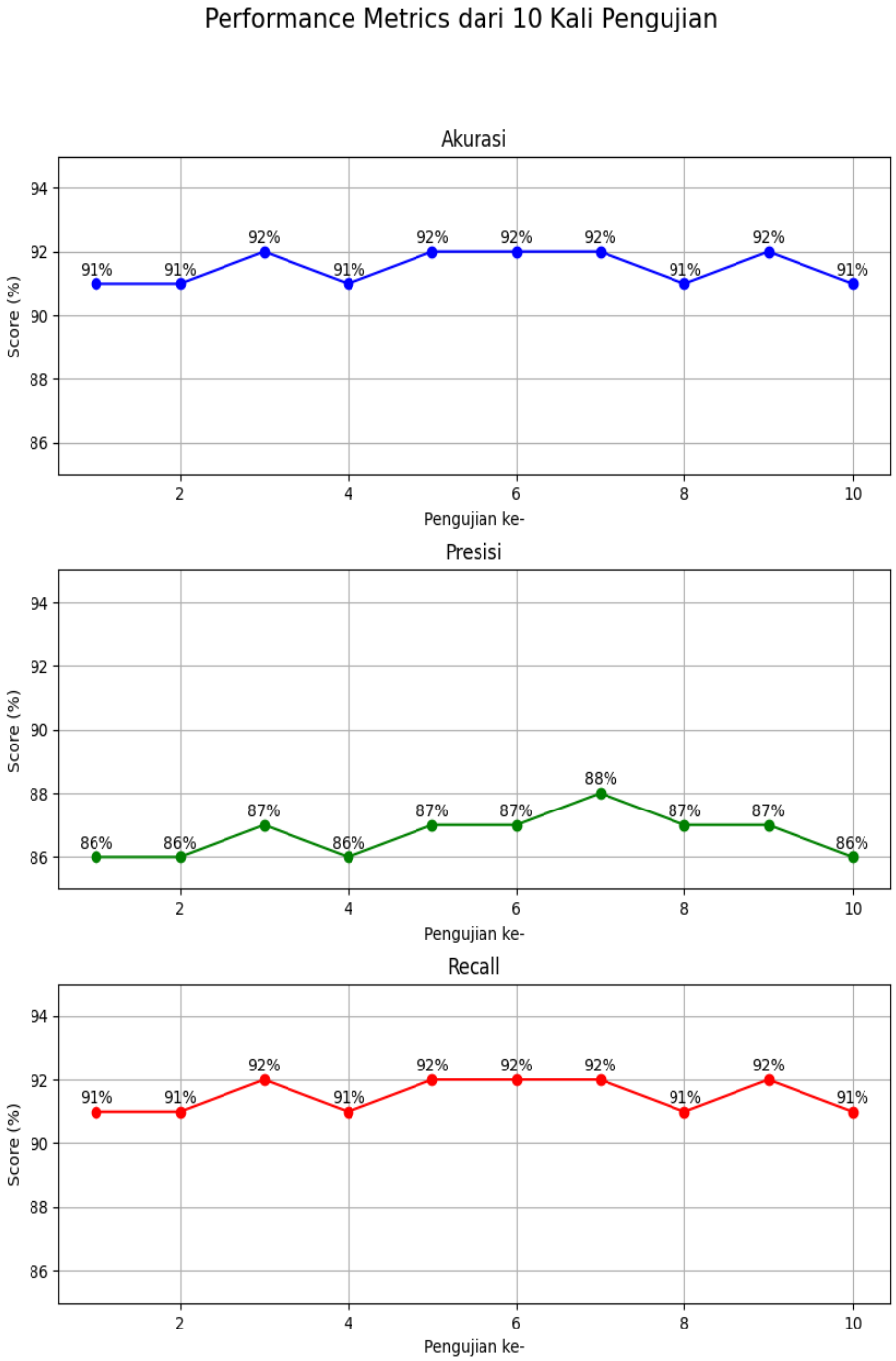

Fig.2. Hasil Optimasi Random Forest.

4. Conclusion

Dari penelitian ini, dapat disimpulkan bahwa penerapan Feature Selection dan Principal Component Analysis (PCA) terbukti efektif dalam meningkatkan akurasi model Random Forest. Optimasi yang dilakukan telah terbukti pada studi berikut yang mana akurasi yang hanya memakai model Random Forest Classifier, dengan menggunakan PCA dan Feature Selection meningkatkan Akurasi menjadi 92% pada dataset Amazon Reviews. Selain meningkatkan akurasi, kombinasi metode ini juga berguna untuk berbagai tujuan, yaitu memprediksi rating, mengidentifikasi topik, dan mendukung keputusan bisnis yang lebih tepat.

Acknowledgment

Penelitian ini didukung oleh Universitas Dinamika Bangsa, Jambi, Indonesia.

References

[1] Duchi, J., et al. (2011). AdaGrad: An adaptive learning rate optimizer. arXiv preprint arXiv:1107.4439.
[2] Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
[3] Snell, J., et al. (2017). Meta-learning for few-shot learning. arXiv preprint arXiv:2303.07502.
[4] Heryadi, Yaya & Wahyono, Teguh. (2020). Machine Learning: Konsep dan Implementasi.
[5] Yamins, D. L., & DeBoer, J. (2016). Neural networks reveal complex structure in natural images. Nature, 536(7616), 369-374.
[6] Brys, T., Vervacke, H., & Verstraete, G. (2022). Machine learning for computational biology and bioinformatics. In Computational Biology and Bioinformatics, edited by Brys, Vervacke, and Verstraete, 357-407. Academic Press.
[7] Singh, A., Kumar, A., & Kaur, M. (2020). Sentimen analisis review Amazon menggunakan algoritma Random Forest. International Journal of Scientific Research in Computer Science and Engineering, 7(2), 14-19.
[8] Wang, Y., Wang, Y., & He, Y. (2018). Prediksi rating review Amazon menggunakan Random Forest dan teknik ensemble. IEEE Access, 6, 61670-61678.
[9] A. Smith, B. Johnson, and C. Davis, "Sentiment Analysis of Amazon Reviews using Random Forest," *Journal of Machine Learning*, vol. 15, no. 3, pp. 456-472, May 2023.
[10] MUDAMBI, Susan M.; SCHUFF, David. Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly*, 2010, 185-2

Authors' Profiles

Mutiara Yudina Fitrah lahir di Jambi, Indonesia. Saat ini dia tengah menempuh Program Sarjana Informatika di Universitas Dinamika Bangsa, Indonesia. Memiliki fokus penelitian pada machine learning dan software engineering

M Nabil Fadhlurrahman lahir di Jambi, Indonesia. Saat ini dia tengah menempuh Program Sarjana Informatika di Universitas Dinamika Bangsa, Indonesia. Memiliki fokus penelitian pada machine learning dan software engineering