JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 66~73
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Optimasi Strategi Pemasaran *E-Commerce* Melalui Prediksi Konversi Berbasis *Machine Learning*

[1]Agustina Heryati, [2]Terttiaavini, [3]Septa Cahyani, [4]K.Ghazali, [5]Harsi Romli, [6]Iski Zaliman
[1,2,3,5]Universitas Indo Global Mandiri, Indonesia
[4]Poltek Prasetiya Mandiri, Indonesia
[6]Universitas Bangka Belitung, Indonesia
[1]agustina.heryati@uigm.ac.id; [2]avini.saputra@uigm.ac.id; [3]septacahyani@uigm.ac.id; [4]igoup95@gmail.com; [5]harsi_romli@uigm.ac.id; [6]iski.zaliman@ubb.ac.id;

Article Info

Article history:

Received, 2024-12-23
Revised, 2024-12-24
Accepted, 2024-12-28

Kata Kunci:

Klasterisasi
Klasifikasi
K-Means clustering
Random Forest
Silhouette Coefficient
TikTok

ABSTRAK

Penelitian ini mengidentifikasi permasalahan dalam meningkatkan konversi penjualan *e-commerce* melalui TikTok, di tengah persaingan konten yang ketat. Tujuan penelitian adalah mengembangkan strategi pemasaran berbasis machine learning untuk menganalisis perilaku pengguna dan mengelompokkan mereka menjadi *Non-Purchasers* dan *Purchasers*. Metode yang digunakan meliputi klasterisasi dengan algoritma *K-Means, K-Medoids,* dan *Fuzzy C-Means*, di mana *K-Means* menunjukkan performa terbaik dengan *Silhouette Coefficient* tertinggi (0.1857) dan *Davies-Bouldin Index* terendah (1.9991). Setelah klasterisasi, dilakukan klasifikasi menggunakan *Naïve Bayes, Decision Tree,* dan *Random Forest*. Model *Random Forest* memberikan hasil terbaik dengan akurasi 0.9945, menunjukkan kemampuannya dalam memprediksi konversi penjualan secara efektif. Hasil dari penelitian ini adalah bahwa *K-Means* dan *Random Forest* adalah metode terbaik dalam klasterisasi dan klasifikasi, masing-masing, untuk memahami perilaku pengguna di TikTok. Temuan ini dapat membantu pelaku *e-commerce* dalam menyesuaikan strategi pemasaran mereka, meningkatkan konversi penjualan, serta efisiensi pengeluaran iklan

Keywords:

Clusterization
Classification
K-Means clustering
Random Forest
Silhouette Coefficient
TikTok

ABSTRACT

The research identifies the problem of enhancing e-commerce sales conversion through TikTok amidst intense content competition. The objective of the study is to develop a machine learning-based marketing strategy to analyze user behavior and categorize them into Non-Purchasers and Purchasers.The method employed includes clustering using K-Means, K-Medoids, and Fuzzy C-Means algorithms, with K-Means demonstrating the best performance, achieving the highest Silhouette Coefficient (0.1857) and the lowest Davies-Bouldin Index (1.9991). Following clustering, classification is performed using Naïve Bayes, Decision Tree, and Random Forest algorithms. The Random Forest model yields the best results with an accuracy of 0.9945, showcasing its effectiveness in predicting sales conversions.The conclusion of this study indicates that K-Means and Random Forest are the optimal methods for clustering and classification, respectively, in understanding user behavior on TikTok. These findings can assist e-commerce players in tailoring their marketing strategies, improving sales conversion rates, and enhancing advertising efficiency.

Penulis Korespondensi:

Agustina Heryati
Program Sistem Informatika,
Universitas Indo Global Mandiri,
Email: agustina.heryati@uigm.ac.id

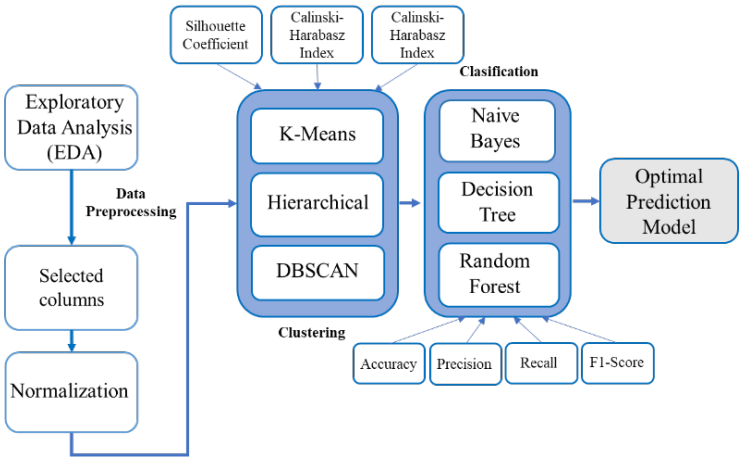
JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 66~73
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

1. PENDAHULUAN

Perkembangan pesat sektor *e-commerce* dalam beberapa tahun terakhir telah menciptakan peluang besar bagi perusahaan untuk meningkatkan jangkauan pasar dan kinerja penjualan mereka. Dalam menghadapi kompetisi yang semakin ketat, penting bagi perusahaan untuk memiliki strategi pemasaran yang efektif agar dapat menarik pelanggan potensial dan meningkatkan tingkat konversi dari pengunjung menjadi pembeli. Salah satu tantangan utama dalam pemasaran digital adalah memahami perilaku pelanggan yang sangat dinamis dan memprediksi kemungkinan konversi dari aktivitas yang dilakukan oleh pengguna di *platform e-commerce*.
Tiktok, sebagai salah satu platform media sosial yang paling populer di dunia, telah menjadi alat pemasaran yang sangat efektif bagi berbagai bisnis, termasuk *e-commerce*. Dengan lebih dari miliaran pengguna aktif bulanan, Tiktok menawarkan peluang besar untuk menjangkau audiens yang luas dan beragam. Namun, popularitas ini juga menghadirkan tantangan bagi pelaku bisnis dalam memaksimalkan potensi *platform* untuk meningkatkan penjualan [1].
Salah satu permasalahan utama adalah tingginya persaingan konten untuk menarik perhatian pengguna. Hal ini membuat bisnis kesulitan memastikan bahwa konten mereka dapat mencapai audiens yang relevan dan memberikan dampak signifikan terhadap konversi penjualan. Dalam konteks *e-commerce*, konversi mengacu pada tindakan spesifik seperti pembelian produk setelah melihat promosi di Tiktok. Perilaku konsumen yang dinamis dan algoritma rekomendasi Tiktok yang kompleks semakin menambah tantangan dalam memahami preferensi pengguna serta mengoptimalkan strategi pemasaran [2]
Penelitian ini bertujuan untuk mengembangkan strategi pemasaran berbasis *machine learning* yang dapat membantu *e-commerce* dalam mengoptimalkan penjualan melalui Tiktok [3]. Secara khusus, penelitian ini bertujuan untuk menganalisis perilaku pengguna dengan membagi data pengguna menjadi dua kelompok utama, yaitu *Non-Purchasers* dan *Purchasers*, melalui klasterisasi [4], [5].
Penelitian ini menerapkan teknik *hybrid* models, di mana data yang digunakan berasal dari data penjualan di Tiktok pada suatu toko online dengan total 1000 data. Untuk menghasilkan klaster *Non-Purchasers* dan *Purchasers*, dilakukan teknik *clustering* menggunakan algoritma *K-Means, K-Medoids,* dan *Fuzzy C-Means* [6], [7]. Selanjutnya, untuk mengevaluasi hasil klasterisasi, digunakan metrik evaluasi yaitu *Silhouette Coefficient* [8], *Calinski-Harabasz Index* [9], dan *Davies-Bouldin Index* [10] untuk mendapatkan hasil klasterisasi yang terbaik. Setelah itu, dilakukan teknik klasifikasi menggunakan metode *Naïve Bayes* dan *Decision Tree*, dengan membandingkan nilai *Accuracy, Precision, Recall,* dan *F1-Score* untuk mengukur hasil klasifikasi yang paling optimal [5], [11]. Untuk menghindari *overfitting*, dilakukan pengujian menggunakan *Cross-Validation*. Berdasarkan hasil evaluasi ini, model prediksi yang paling optimal [12].
Penelitian ini diharapkan dapat memberikan wawasan bagi pelaku *e-commerce* di Tiktok dalam mengoptimalkan strategi pemasaran mereka dengan memahami perbedaan perilaku antara *Purchasers* dan *Non-Purchasers*, yang dapat membantu suatu toko online untuk menyesuaikan konten dan promosi secara lebih tepat sasaran, sehingga dapat meningkatkan konversi penjualan, memperbaiki pengalaman pengguna, dan efisiensi pengeluaran iklan. Dengan memahami karakteristik masing-masing kelompok, toko online dapat lebih efektif dalam menargetkan audiens yang memiliki potensi untuk melakukan pembelian, serta meningkatkan interaksi dan keterlibatan pengguna di platform Tiktok.

2. METODE PENELITIAN

Metodologi penelitian ini menjelaskan langkah-langkah sistematis yang digunakan untuk mengumpulkan, menganalisis, dan menginterpretasikan data guna mencapai tujuan penelitian yang telah ditetapkan. Gambar 1 merupakan *Machine Learning Workflow*.



Gambar 1. *Machine Learning Workflow*

Gambar 1 di atas menunjukkan *workflow machine learning* yang mencakup tahapan utama dalam pengolahan data dan pengembangan model prediksi. Proses dimulai dari data *preprocessing*, yang bertujuan untuk mempersiapkan data agar siap digunakan dalam analisis lebih lanjut. Selanjutnya, tahapan modeling dilakukan, yang terdiri dari dua bagian utama: klasterisasi menggunakan metode seperti *K-Means, K-Medoids,* dan *Fuzzy C-Means*, dan klasifikasi menggunakan algoritme *Naive Bayes, Decision Tree,* atau *Random Forest.* Setelah itu, dilakukan evaluasi untuk memilih model prediksi yang optimal, yang mampu memberikan hasil terbaik sesuai dengan metrik evaluasi yang ditentukan. Alur ini menggambarkan pendekatan sistematis dalam pengembangan solusi berbasis machine learning

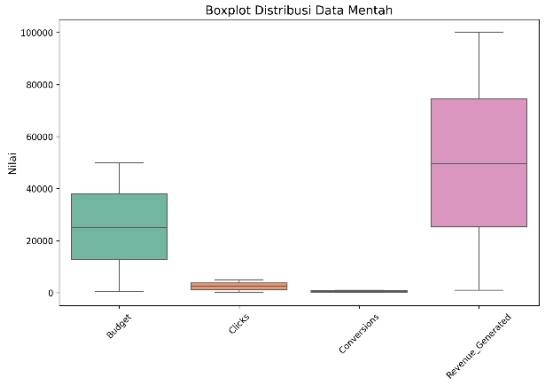
***Exploratory Data Analysis* (EDA)**

Langkah awal dalam proses analisis data yang bertujuan untuk memahami karakteristik, pola, dan hubungan antar variabel dalam dataset. Dalam penelitian ini, EDA digunakan untuk mengeksplorasi data terkait kinerja pemasaran dan produk, sehingga dapat memberikan gambaran mendalam mengenai distribusi, korelasi antar fitur, potensi *outlier*, serta pola yang relevan untuk mendukung analisis lebih lanjut. Proses ini juga membantu dalam mendeteksi anomali atau inkonsistensi yang mungkin memengaruhi hasil penelitian.
Analisa data dilakukan dengan menggunakan deskripsi statistik seperti mean, median, standar deviasi, serta distribusi frekuensi untuk setiap atribut utama, yaitu *'Budget'*, *'Clicks'*, *'Conversions'*, dan *'Revenue_Generated'*. Hasil analisis deskriptif disajikan pada Tabel 1.

Tabel 1. Deskripsi statistic data Tiktok

Atribut	*Budget*	*Clicks*	*Conversions*	*Revenue_Generated*
mean	25263.60	2481.90	498.97	50038.62
std	14350.08	1435.97	289.47	28545.70
min	500.44	10.00	1.00	1002.08
25%	12789.19	1225.75	247.00	25264.25
50%	25030.17	2451.00	499.00	49513.81
75%	37921.72	3723.00	751.00	74507.15
max	49999.63	4999.00	999.00	99999.47

Selain itu, dilakukan visualisasi data menggunakan boxplot untuk mendeteksi outlier, serta mengidentifikasi hubungan potensial antar variabel. Langkah ini bertujuan untuk memperoleh wawasan awal yang dapat digunakan sebagai dasar untuk tahap analisis dan pemodelan selanjutnya. Gambar 1 merupakan gambar digram *boxplot*.



Gambar 1 Diagram *bloxplot* data penjualan di Tiktok

Berdasarkan *boxplot* yang disajikan, terlihat adanya variasi yang cukup signifikan pada keempat variabel *Budget, Clicks, Conversions,* dan *Revenue Generated*. Variabel *Revenue Generated* memiliki rentang nilai yang paling luas, mengindikasikan adanya perbedaan yang besar dalam pendapatan yang dihasilkan. Terdapat beberapa *outlier*, terutama pada variabel *Conversions*, yang menunjukkan adanya data yang sangat berbeda dari data lainnya. Ini mungkin mengindikasikan adanya beberapa kampanye yang performanya jauh di bawah rata-rata. Secara keseluruhan, *boxplot* ini memberikan gambaran awal bahwa kinerja kampanye atau aktivitas yang diukur cukup beragam

Data *Preprocessing*

Data *preprocessing* dilakukan untuk mempersiapkan data agar siap digunakan dalam proses analisis dan pemodelan [13]. Data diperoleh dari database tiktok pada suatu took online dengan jumlah 1000 data, yang kemudian dipilih kolom/atribut tertentu, yaitu *'Budget'*, *'Clicks'*, *'Conversions'*, dan *'Revenue_Generated'*. Pemilihan atribut ini dilakukan karena dianggap relevan dalam mendukung tujuan penelitian.

JSAI: Journal Scientific and Applied Informatics
Vol. 8, No. 1, Januari 2025, hal. 66~73
E-ISSN: 2614-3054; P-ISSN: 2614-3062, accredited by Kemenristekdikti, Sinta 4
DOI: 10.36085

Berdasarkan hasil struktur data dapat disimpulkan bahwa data tidak memiliki nilai yang kosong atau null pada setiap kolomnya. Hal ini menunjukkan bahwa dataset telah bersih dari missing values, sehingga tidak diperlukan langkah imputasi data.
Untuk menangani *outlier*, digunakan metode *Interquartile Range* (IQR). IQR mengukur rentang antara kuartil pertama (Q1) dan kuartil ketiga (Q3) dalam distribusi data, sehingga memberikan gambaran mengenai sebaran nilai tengah data. Outlier didefinisikan sebagai nilai yang berada di luar rentang Q1−1.5×IQR hingga Q3+1.5×IQR. Data yang terdeteksi sebagai outlier ini kemudian dapat dihapus atau disesuaikan untuk memastikan bahwa analisis selanjutnya tidak dipengaruhi oleh nilai ekstrem yang dapat mendistorsi hasil analisis dan model prediksi. Selanjutnya, dilakukan normalisasi data menggunakan metode StandardScaler untuk menyelaraskan skala nilai dari setiap atribut [14]. Hal ini diperlukan karena atribut memiliki rentang skala yang berbeda, seperti *'Budget'* yang dinyatakan dalam satuan puluhan ribu, sementara *'Conversions'* memiliki nilai dalam satuan ratusan ribu. Normalisasi ini memastikan bahwa semua atribut memiliki bobot yang seimbang dalam analisis, sehingga algoritme machine learning dapat bekerja secara optimal.

Pengelompokan Data (*Clustering*)

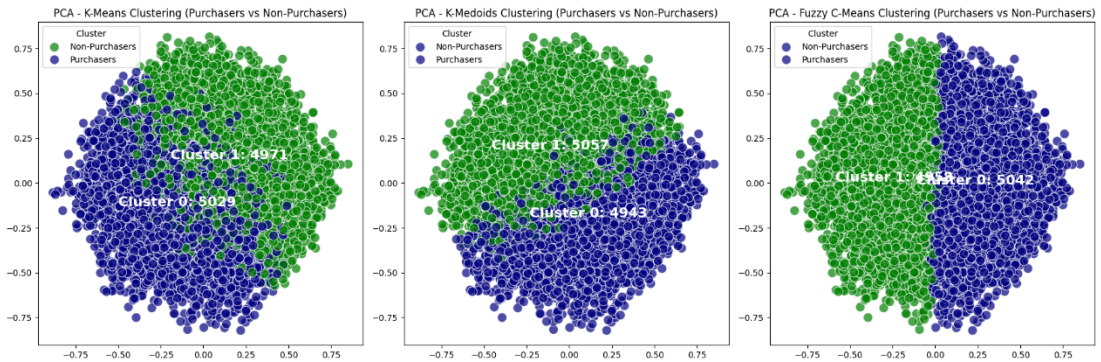
Tahapan *clustering* dalam penelitian ini bertujuan untuk mengelompokkan pengguna berdasarkan perilaku mereka dalam melakukan pembelian atau tidak [4], [15], dengan menggunakan variabel-variabel seperti *Budget, Clicks, Conversions,* dan *Revenue Generated* yang menggambarkan interaksi pengguna dengan kampanye pemasaran. Teknik clustering diterapkan untuk mengelompokkan pengguna dengan pola perilaku serupa, sehingga memungkinkan identifikasi segmen pengguna yang lebih cenderung melakukan pembelian dan yang tidak [1], [16]. Hal ini mendukung penyusunan strategi pemasaran yang lebih efektif dengan menargetkan segmen-segmen dengan potensi konversi lebih tinggi. Hasil dari proses *clustering* kemudian dievaluasi menggunakan tiga metrik utama, yaitu *Silhouette Coefficient, Calinski-Harabasz Index,* dan *Davies-Bouldin Index*, yang masing-masing memberikan gambaran tentang kualitas klaster yang terbentuk dari setiap algoritma *clustering* yang digunakan. Tabel 2. Menampilkan hasil klasterisasi tersebut :

Tabel 2. Metrik Evaluasi Performa Metode *Clustering*

Metode Klasterisasi	*Silhouette Coefficient*	*Calinski-Harabasz Index*	*Davies-Bouldin Index*
K-Means	0.1857	2336.4650	1.9991
KMedoids	0.1663	2003.0720	2.1508
Fuzzy C-Means	0.1701	2070.9241	2.1144

Berdasarkan hasil evaluasi menggunakan tiga metrik, *K-Means* menunjukkan performa terbaik dalam hal pemisahan klaster. Pada *Silhouette Score*, *K-Means* memperoleh nilai tertinggi (0.1857), yang menunjukkan pemisahan klaster yang lebih baik dibandingkan dengan *K-Medoids* (0.1663) dan *Fuzzy C-Means* (0.1701). Begitu pula pada *Calinski-Harabasz Index, K-Means* (2336.4650) menunjukkan klaster yang lebih terpisah dan kohesif, sementara *K-Medoids* (2003.0720) dan *Fuzzy C-Means* (2070.9241) menghasilkan klaster yang kurang *koheren*. Terakhir, pada *Davies-Bouldin Index, K-Means* memiliki nilai terendah (1.9991), yang mengindikasikan pemisahan klaster yang lebih jelas dan sedikit tumpang tindih dibandingkan dengan *K-Medoids* (2.1508) dan *Fuzzy C-Means* (2.1144). Secara keseluruhan, *K-Means* memberikan pemisahan klaster yang lebih efektif dengan lebih sedikit tumpang tindih dan koherensi yang lebih baik dibandingkan dengan kedua metode lainnya.

Untuk memberikan gambaran yang jelas dan intuitif mengenai bagaimana masing-masing algoritma membagi data ke dalam klaster. Dengan menggunakan scatter plot, dapat dengan mudah melihat pemisahan antar klaster, serta perbedaan dalam ketajaman dan kejelasan batas antar klaster yang dihasilkan oleh ketiga metode tersebut. Gambar 2 merupakan visualisasi dari tiga metode tersebut menggunakan *scater* plot.



Gambar 2. Visualisasi hasil dari *clusterisasi* menggunakan *scater plot*

Visualisasi hasil klastering yang ditampilkan menunjukkan perbandingan performa tiga algoritma klastering yang diterapkan pada data yang telah melalui reduksi dimensi menggunakan *Principal Component Analysis* (PCA) [17]. Pada plot *PCA K-Means Clustering*, data terbagi menjadi dua kelompok yang cukup jelas, meskipun ada beberapa titik yang sulit diklasifikasikan dengan pasti. *PCA K-Medoids Clustering* menunjukkan pemisahan yang sedikit lebih baik, terutama di area perbatasan antar kelompok, berkat kemampuannya yang lebih *robust* terhadap *outlier*. Sedangkan pada plot *PCA Fuzzy C-Means Clustering,* batas antara kedua kelompok terlihat lebih lembut, mencerminkan sifat *fuzzy* dari algoritma ini yang memungkinkan suatu titik data menjadi anggota dari beberapa klaster sekaligus dengan tingkat keanggotaan yang berbeda. Secara keseluruhan, ketiga algoritma ini memberikan hasil yang berbeda dalam hal ketajaman pemisahan antar kelompok, dengan *K-Means* menunjukkan pemisahan yang cukup jelas, *K-Medoids* lebih baik di area perbatasan, dan *Fuzzy C-Means* menghasilkan batas yang lebih fleksibel antara klaster.

Klasifikasi (*Clasification*)

Setelah melakukan clusterisasi dan mendapatkan *cluster* terbaik yaitu *K-Means Clustering*, langkah selanjutnya adalah melakukan klasifikasi terhadap hasil *cluster* tersebut untuk memprediksi konversi pembelian. Algoritma yang digunakan untuk klasifikasi ini adalah *Naïve Bayes, Decision Tree,* dan *Random Forest*. Setiap algoritma diuji untuk menentukan model mana yang paling efektif dalam mengklasifikasikan pengguna berdasarkan *cluster* yang telah terbentuk, dengan tujuan untuk memprediksi apakah pengguna tersebut berpotensi melakukan pembelian atau tidak. Kinerja masing-masing model dievaluasi menggunakan metrik akurasi, *precision*, *recall*, dan *F1-Score* untuk memilih model terbaik yang dapat memberikan prediksi yang akurat dan dapat diandalkan dalam mendukung strategi pemasaran yang lebih terarah [18]. Hasil dari pengujian *clustering* memperoleh nilai yang ditampilkan pada Tabel 1.

Tabel 3. Hasil klasifikasi menggunakan metrik evaluasi

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Naïve Bayes	0.9875	0.9875	0.9875	0.9875
Decision Tree	0.9925	0.9925	0.9925	0.9925
Random Forest	0.9945	0.9945	0.9945	0.9945

Hasil pengujian tersebut memperoleh hasil

1. *Naïve Bayes* : Model *Naïve Bayes* menunjukkan performa yang sangat baik dengan akurasi, *precision*, *recall*, dan *F1-Score* semuanya mencapai nilai 0.9875. Hal ini mengindikasikan bahwa model ini mampu mengklasifikasikan data dengan sangat baik, memberikan prediksi yang akurat dan seimbang antara presisi dan recall. Precision yang tinggi menandakan bahwa sebagian besar prediksi positif yang dibuat oleh model benar-benar merupakan positif, sementara recall yang tinggi menunjukkan bahwa sebagian besar data positif terdeteksi dengan benar.
2. *Decision Tree* : Model *Decision Tree* juga memberikan hasil yang sangat baik dengan nilai akurasi, *precision*, *recall*, dan *F1-Score* mencapai 0.9925. Ini menunjukkan bahwa model ini sangat baik dalam memisahkan kelas-kelas yang ada, dengan sedikit kesalahan dalam mengklasifikasikan data uji. *Precision* dan *recall* yang sangat tinggi menunjukkan bahwa model ini sangat efektif dalam mendeteksi positif dan menghindari kesalahan klasifikasi, meskipun ada kemungkinan model ini sedikit lebih rentan terhadap overfitting mengingat perbedaan antara akurasi pelatihan dan pengujian [19].
3. *Random Forest* : Model *Random Forest* memberikan hasil terbaik dengan akurasi, *precision*, *recall*, dan *F1-Score* yang mencapai nilai 0.9945. Keunggulan dari *Random Forest* terlihat dalam kemampuan model ini untuk menggabungkan hasil dari beberapa pohon keputusan, yang memungkinkan model untuk meminimalkan risiko *overfitting* dan lebih baik dalam menggeneralisasi data uji. *Precision* dan *recall* yang tinggi menunjukkan bahwa model ini mampu mendeteksi hampir semua data positif dengan sangat baik dan mengurangi kesalahan klasifikasi.

Perhitungan *Overfitting* dan *Cross-Validation*

Untuk memastikan bahwa model tidak mengalami *overfitting*, maka dilakukan *cross-validation* untuk menguji kestabilan dan generalisasi model pada berbagai subset data [20]. Dengan *cross-validation*, data dibagi menjadi beberapa bagian (*folds*), dan model dilatih serta diuji secara bergantian pada masing-masing bagian. Hal ini memberikan gambaran yang lebih akurat tentang bagaimana model akan bekerja pada data yang belum pernah dilihat sebelumnya. hasil perhitungan *cross-validation* di tampilkan pada tabel 4 berikut ini

Tabel 4. Hasil *Cross-Validation* untuk Model Klasifikasi

Model	***Validation Accuracy***
Naïve Bayes	0.9854
Decision Tree	0.9910
Random Forest	0.9947

Tabel ini menunjukkan hasil akurasi yang diperoleh dengan menggunakan *cross-validation* untuk setiap model. Hasil tersebut mengindikasikan kestabilan dan keandalan model dalam memprediksi data yang tidak terlihat sebelumnya, serta memberikan gambaran bahwa model mampu bekerja dengan baik pada berbagai subset data. Hasil *cross-validation* memberikan hasil yang lebih konsisten dan dapat mengurangi bias dari data yang hanya terbatas pada pelatihan atau pengujian saja.

Dari hasil evaluasi model yang dilakukan, dapat menyimpulkan beberapa hal terkait performa masing-masing model serta kemungkinan terjadinya overfitting.

1. *Naïve Bayes* menunjukkan akurasi *cross-validation* sebesar 0.9854 ± 0.0027, yang hampir setara dengan akurasi pelatihan dan pengujian, mengindikasikan bahwa model ini tidak mengalami *overfitting*. Ini berarti bahwa *Naïve Bayes* mampu menjaga performa yang stabil di seluruh data yang berbeda dan tetap menghasilkan prediksi yang akurat, baik pada data pelatihan maupun data pengujian.
2. *Decision Tree* memperoleh akurasi *cross-validation* sebesar 0.9910 ± 0.0021. Meskipun masih cukup baik, nilai ini sedikit lebih rendah dibandingkan dengan akurasi pelatihan yang mencapai 1.0000, yang semakin menguatkan indikasi bahwa model ini *overfit*. *Cross-validation* ini menunjukkan bahwa meskipun model sangat cocok dengan data pelatihan, performanya pada data yang belum pernah dilihat sedikit menurun.
3. *Random Forest* memperoleh akurasi *cross-validation* sebesar 0.9947 ± 0.0009, yang hampir setara dengan akurasi pengujian dan hanya sedikit lebih rendah dibandingkan dengan akurasi pelatihan. Nilai ini menunjukkan bahwa *Random Forest* lebih mampu mengatasi overfitting dibandingkan dengan *Decision Tree* karena sifat ensemble-nya yang lebih mengandalkan agregasi dari berbagai pohon keputusan, memberikan model yang lebih stabil.

3. HASIL DAN ANALISIS

Interpretasi Hasil dapat diartikan sebagai langkah di mana peneliti menganalisis dan memberikan penjelasan mengenai hasil yang diperoleh dari eksperimen atau analisis data yang telah dilakukan. Pada bagian ini, hasil-hasil yang didapatkan dari proses seperti clustering dan klasifikasi dijelaskan dengan cara yang mendalam, sehingga memberikan pemahaman yang jelas tentang arti dari angka-angka atau metrik yang diperoleh dan implikasi dari temuan tersebut.

1. Pada tahap *clustering*, tujuan utama adalah mengelompokkan pengguna berdasarkan perilaku mereka dalam melakukan pembelian atau tidak. Berdasarkan hasil evaluasi *clustering*, *K-Means* terbukti menjadi metode yang paling optimal. *K-Means* memperoleh nilai *Silhouette Coefficient* tertinggi, yaitu 0.1857, yang menunjukkan bahwa klaster yang terbentuk memiliki pemisahan yang jelas antar klaster, dengan klaster-klaster yang lebih terpisah dengan baik. Di sisi lain, *Calinski-Harabasz Index K-Means* yang tinggi, 2336.4650, menunjukkan kohesi klaster yang baik dan pemisahan yang jelas antar klaster, sementara *Davies-Bouldin Index* terendah, yaitu 1.9991, menunjukkan sedikit tumpang tindih antar klaster. Sebaliknya, meskipun *K-Medoids* dan *Fuzzy C-Means* memberikan hasil yang cukup baik, kedua metode ini menghasilkan klaster yang lebih kurang koheren, dengan nilai *Silhouette Coefficient* yang lebih rendah dan *Davies-Bouldin Index* yang lebih tinggi, yang berarti pemisahan antar klaster tidak sebaik *K-Means*. Oleh karena itu, *K-Means* dipilih sebagai metode *clustering* yang paling optimal untuk penelitian ini karena pemisahan klaster yang lebih baik dan lebih efektif dalam mengidentifikasi segmen pengguna berdasarkan perilaku pembelian.
2. Pada tahap klasifikasi, model digunakan untuk memprediksi apakah pengguna yang telah dikelompokkan menjadi *Non-Purchasers* dan *Purchasers*. Berdasarkan hasil evaluasi menggunakan metrik *Accuracy*, *Precision, Recall,* dan *F1-Score*, model *Random Forest* menunjukkan performa terbaik, dengan nilai akurasinya mencapai 0.9945. Model ini juga menghasilkan nilai *precision, recall,* dan *F1-Score* yang sangat tinggi, yang menunjukkan kemampuannya untuk mendeteksi hampir semua data positif dengan sangat baik dan mengurangi kesalahan klasifikasi. *Decision Tree* juga menunjukkan hasil yang sangat baik dengan akurasi 0.9925, namun sedikit lebih rentan terhadap *overfitting*, yang mengindikasikan perbedaan antara performa pada data pelatihan dan pengujian. Sementara itu, *Naïve Bayes* memiliki akurasi 0.9875, yang menunjukkan prediksi yang sangat akurat dan seimbang antara *precision* dan *recall*, meskipun performanya sedikit lebih rendah dibandingkan dengan *Decision Tree* dan *Random Forest*. Dengan demikian, *Random Forest* dipilih sebagai model klasifikasi yang paling unggul dalam memprediksi konversi pembelian dalam penelitian ini, karena memberikan hasil terbaik pada semua metrik evaluasi.
3. Untuk memastikan model tidak mengalami *overfitting*, dilakukan *cross-validation* yang membagi data menjadi beberapa bagian untuk menguji kestabilan dan generalisasi model pada data yang tidak terlihat sebelumnya. Berdasarkan hasil *cross-validation,* model *Naïve Bayes* memperoleh akurasi 0.9854, yang hampir setara dengan akurasi pelatihan dan pengujian, mengindikasikan bahwa model ini tidak mengalami *overfitting* dan mampu memberikan prediksi yang stabil pada berbagai subset data. Sementara itu, *Decision Tree* menghasilkan akurasi 0.9910, yang sedikit lebih rendah dibandingkan akurasi pelatihannya yang mencapai 1.0000, menunjukkan adanya sedikit *overfitting*. Sebaliknya, *Random Forest* menghasilkan akurasi 0.9947, yang hampir setara dengan akurasi pengujian dan sedikit lebih rendah

dibandingkan akurasi pelatihan, menunjukkan bahwa model ini lebih stabil dan dapat menghindari *overfitting* lebih baik dibandingkan dengan *Decision Tree*. Dengan demikian, *Random Forest* terbukti lebih andal dalam memprediksi pembelian pengguna pada data yang baru dan lebih mampu mengatasi overfitting dibandingkan dengan *Decision Tree*.

4. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa metode *K-Means* merupakan algoritma *clustering* yang paling optimal untuk mengelompokkan pengguna berdasarkan perilaku pembelian mereka. Metode ini menunjukkan pemisahan klaster yang jelas dengan nilai *Silhouette Coefficient* tertinggi dan nilai *Calinski-Harabasz Index* yang mencerminkan kohesi klaster yang baik serta pemisahan antar klaster yang optimal. Meskipun algoritma *K-Medoids* dan *Fuzzy C-Means* juga memberikan hasil yang cukup baik, *K-Means* lebih unggul dalam mengidentifikasi segmen pengguna dengan potensi konversi yang lebih tinggi. Keunggulan ini terlihat dari pemisahan klaster yang lebih jelas dan kohesif, dengan hasil klasterisasi menunjukkan *Non-Purchasers* (*Cluster 1*) = 4.971 dan *Purchasers* (*Cluster 2*) = 5.029. Pada tahap klasifikasi, model *Random Forest* menunjukkan performa terbaik dalam memprediksi konversi pembelian pengguna. Model ini memiliki nilai akurasi, *precision*, *recall*, dan *F1-Score* tertinggi, yang menandakan kemampuannya untuk mendeteksi hampir semua data positif dengan sangat baik serta meminimalkan kesalahan klasifikasi. *Decision Tree* dan *Naïve Bayes* juga memberikan performa yang baik, tetapi *Random Forest* lebih unggul dalam hal stabilitas dan generalisasi, terutama dalam menghindari *overfitting*.

REFERENSI

[1] C. Wusylko *et al.*, "Using machine learning techniques to investigate learner engagement with TikTok media literacy campaigns," *Journal of Research on Technology in Education*, vol. 56, no. 1, pp. 72–93, 2024, doi: 10.1080/15391523.2023.2266518.
[2] Y. Nie and Y. Xu, "Prediction On Tiktok Like Behavior Based on Random Forest Model," 2024.
[3] M. Sihag *et al.*, "A Data-Driven Approach for Finding Requirements Relevant Feedback from TikTok and YouTube," *Proceedings of the IEEE International Conference on Requirements Engineering*, vol. 2023-September, pp. 111–122, 2023, doi: 10.1109/RE57278.2023.00020.
[4] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.
[5] "Supplementary Material to Classified Mixed Model Prediction."
[6] A. Avram, O. Matei, C.-M. Pintea, P. C. Pop, and C. A. Anton, "Comparative Analysis of Clustering Techniques for a Hybrid Model Implementation," in *15th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2020)*, Á. Herrero, C. Cambra, D. Urda, J. JSedano, H. Quintián, and E. Corchado, Eds., Springer, Cham, 2021, pp. 22–32. doi: 10.1007/978-3-030-57802-2_3.
[7] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review," *Journal of Data Analysis and Information Processing*, vol. 08, no. 04, pp. 341–357, 2020, doi: 10.4236/jdaip.2020.84020.
[8] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, pp. 747–748, Oct. 2020, doi: 10.1109/DSAA49011.2020.00096.
[9] S. P. Lima and M. D. Cruz, "A genetic algorithm using Calinski-Harabasz index for automatic clustering problem," *Revista Brasileira de Computação Aplicada*, vol. 12, no. 3, pp. 97–106, Sep. 2020, doi: 10.5335/rbca.v12i3.11117.
[10] M. Mughnyanti, S. Efendi, and M. Zarlis, "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jan. 2020. doi: 10.1088/1757-899X/725/1/012128.
[11] . T. *et al.*, "Clustering Analysis of Premier Research Fields," *International Journal of Engineering & Technology*, vol. 7, no. 4.44, 2018, doi: 10.14419/ijet.v7i4.44.26860.
[12] M. A. Salam, A. T. Azar, M. S. Elgendy, and K. M. Fouad, "The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, pp. 641–655, 2021, doi: 10.14569/IJACSA.2021.0120480.
[13] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "Big Data Preprocessing: Enabling Smart Data," *Big Data Preprocessing: Enabling Smart Data*, pp. 1–186, Jan. 2020, doi: 10.1007/978-3-030-39105-8/COVER.
[14] H. Henderi, "Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor (kNN) Algorithm to Test the Accuracy of Types of Breast Cancer," *IJIIS: International*

Journal of Informatics and Information Systems, vol. 4, no. 1, pp. 13–20, Mar. 2021, doi: 10.47738/ijiis.v4i1.73.

[15] D. Marcelina, A. Kurnia, and T. Terttiaavini, "Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 293–301, Nov. 2023, doi: 10.57152/malcom.v3i2.952.

[16] N. Thakur *et al.*, "A Labelled Dataset for Sentiment Analysis of Videos on YouTube, TikTok, and Other Sources about the 2024 Outbreak of Measles," Jun. 2024, doi: 10.21227/40S8-XF63.

[17] M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D'Enza, A. Markos, and E. Tuzhilina, "Principal component analysis," *Nature Reviews Methods Primers 2022 2:1*, vol. 2, no. 1, pp. 1–21, Dec. 2022, doi: 10.1038/s43586-022-00184-w.

[18] Ž. Vujović, "Classification Model Evaluation Metrics," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 599–606, 2021, doi: 10.14569/IJACSA.2021.0120670.

[19] K. Furmańczyk, K. Paczutkowski, M. Dudziński, and D. Dziewa-Dawidczyk, "Classification Methods Based on Fitting Logistic Regression to Positive and Unlabeled Data," in *Computational Science – ICCS 2022: 22nd International Conference, London, UK,* D. Groen, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot, Eds., London: Springer Cham, Jun. 2022, pp. 31–45. doi: 10.1007/978-3-031-08751-6_3.

[20] C. Mahlich, T. Vente, and J. Beel, "From Theory to Practice: Implementing and Evaluating e-Fold Cross-Validation," in *International Conference on Artificial Intelligence and Machine Learning Research (CAIMLR). 2024*, 2024. doi: https://doi.org/10.48550/arXiv.2410.09463.