Vol. 11, No.1, Juni 2026, 11 - 22

Penerapan K-Means dan Algoritma C4.5 dalam Klasifikasi Ulasan Pengguna Aplikasi Mitsubishi SFID

Dzil Hidayati [1,*], Sarjon Defit [2], Billy Hendrik [3]

[1,2,3] Magister Teknik Informatika; Universitas Putra Indonesia "YPTK" Padang; Jl. Raya Lubuk Begalung Padang - Sumatera Barat, Indonesia; e-mail: hidayati.dzil@gmail.com, sarjond@yahoo.co.uk, billy_hendrik@upiyptk.ac.id.

* Korespondensi: e-mail: hidayati.dzil@gmail.com

Diterima: 20 Februari 2026; Review: 25 Februari 2026; Disetujui: 27 Februari 2026

Cara sitasi: Hidayati D, Defit S, Hendrik B. 2026. Penerapan K-Means dan Algoritma C4.5 dalam Klasifikasi Ulasan Pengguna Aplikasi Mitsubishi SFID. Vol 11(1): 11-22.

Abstrak: Penelitian ini bertujuan untuk mengelompokkan dan mengklasifikasikan ulasan pengguna aplikasi Mitsubishi SFID secara sistematis guna menghasilkan informasi yang terstruktur sebagai dasar pengambilan keputusan dalam pengembangan aplikasi. Banyaknya ulasan yang memuat keluhan, kritik, saran, serta penilaian dengan variasi bahasa yang beragam menyebabkan proses analisis secara manual menjadi tidak efisien, memerlukan waktu yang lama, dan berpotensi menimbulkan subjektivitas. Kondisi tersebut mengakibatkan informasi penting terkait pengalaman pengguna, kendala teknis, serta kualitas layanan belum dimanfaatkan secara optimal sebagai bahan evaluasi. Oleh karena itu, diperlukan pendekatan berbasis data mining untuk mengolah dan menganalisis ulasan secara otomatis dan objektif. Metode yang digunakan dalam penelitian ini adalah K-Means Clustering dan algoritma C4.5. K-Means Clustering diterapkan untuk mengelompokkan ulasan berdasarkan tingkat kemiripan karakteristik dan kesamaan tema yang terkandung dalam teks ulasan. Selanjutnya, algoritma C4.5 digunakan untuk membangun model klasifikasi berbasis pohon keputusan dalam mengategorikan ulasan ke dalam kelas tertentu berdasarkan atribut yang relevan. Dataset yang digunakan berasal dari ulasan pengguna aplikasi Mitsubishi SFID yang diperoleh melalui platform Google Play Store dan telah melalui tahapan praproses data, meliputi pembersihan teks, normalisasi, serta pembentukan atribut sebelum dilakukan proses analisis. Hasil penelitian menunjukkan bahwa model klasifikasi yang dibangun mampu mencapai tingkat akurasi sebesar 88,89% dengan performa yang stabil dan konsisten. Informasi yang dihasilkan mampu menggambarkan permasalahan dan kebutuhan pengguna secara lebih terstruktur. Selain memberikan kontribusi praktis bagi pengembangan aplikasi, penelitian ini juga dapat dijadikan sebagai acuan dan referensi bagi penelitian selanjutnya yang membahas analisis ulasan pengguna, penerapan metode clustering dan klasifikasi, serta pengembangan model pengambilan keputusan berbasis data.
Kata kunci: Sentimen Analisis, Ulasan Aplikasi, clustering, K-Means, Algoritma C4.5,

Abstract: *This study aims to systematically cluster and classify user reviews of the Mitsubishi SFID application to generate structured information that supports decision-making in application development. The large number of reviews containing complaints, criticisms, suggestions, and ratings expressed in diverse language variations makes manual analysis inefficient, time-consuming, and potentially subjective. As a result, important information related to user experience, technical issues, and service quality has not been optimally utilized as a basis for evaluation. Therefore, a data mining approach is required to process and analyze user reviews automatically and objectively. The methods applied in this study are K-Means Clustering and the C4.5 algorithm. K-Means Clustering is used to group reviews based on similarity in characteristics and thematic content within the text. Subsequently, the C4.5 algorithm is employed to construct a decision tree-based classification model to categorize reviews into specific classes according to relevant attributes. The dataset consists of user reviews of the Mitsubishi SFID application collected from the Google Play Store platform. The data underwent*

preprocessing stages, including text cleaning, normalization, and attribute construction, prior to clustering and classification. The results indicate that the developed classification model achieved an accuracy rate of 88.89%, demonstrating stable and consistent performance in classifying user reviews. The findings provide structured insights into user problems and needs, which can be utilized to prioritize feature improvements and enhance service quality. In addition to offering practical contributions to application development, this study can serve as a reference for future research related to user review analysis, the implementation of clustering and classification methods, and data-driven decision-making models..

***Keywords**: Sentiment Analysis, Application Review, Clustering, K-means, Algoritma C4.5*

1. Pendahuluan

Industri otomotif Indonesia yang terkemuka diantaranya adalah Toyota, Honda, Suzuki, Mitsubishi dan beberapa merk lainnya yang bisa mengembangkan sayap d indonesia, perusahaan tersebut menjadi pilar penting dalam hal bidang manufaktur serta memiliki peran dalam memajukan Indonesia ke dalam rantai global[1]. Perusahaan Mitsubishi yang ada di Indonesia dioperasikan oleh PT Krama Yudha Tiga Berlian Motors (KTB) yang focus pada bisnis kendaraan niaga dan PT Mitsubishi Motors Krama Yudha Sales Indonesia (MMKSI) yang focus pada pemasaran dan penjualan kendaraan penumpang dan dimana PT tersebut merupakan anak perusahaan dari Mitsubishi Motors Corporation [2]. PT. MMKSI didukung oleh jaringan dealer yang menjadi tiang utama dalam penjualan dan layanan. Dealer-dealer Mitsubishi telah tersebar dan berkembang di seluruh penjuru Indonesia, setiap dealer memiliki fasilitas lengkap untuk mendukung proses penjualan dan layanan kepada pelanggan dimana saat ini MMKSI memiliki 317 dealer resmi yang didukung oleh sekitar 5.100 salesman yang sudah terlatih [3] Aplikasi SFID sebagai media platform informasi dan manajemen aktivitas penjualan yang dilakukan oleh tenaga penjual dealer secara digital [4] Dalam penggunakan aplikasi SFID, aplikasi ini telah menerima berbagai ulasan dari pengguna yaitu berupa apresiasi terhadap fitur maupun keluhan atas kendala teknis yang dialami yang berdampak pada proses penjualan menjadi kurang optimal[5].

Keluhan ulasan dari aplikasi memiliki peran penting dalam proses software engineering terutama pada fase operasional dan peningkatan mutu, analisis terhadap keluhan yang memungkinkan pengembang untuk menemukan bug, mengevaluasi kestabilan versi terbaru, serta mengerti kebutuhan pengguna yang belum terpenuhi [6]. Penelitian lainnya yang memahami pengalaman pengguna bahwa ulasan dapat mencerminkan kepuasan, harapan, serta keluhan pengguna mengenai fitur aplikasi [7]. Pada ulasan aplikasi tantangan lain yang muncul adalah besarnya volume ulasan yang masuk serta keberagaman isi dan gaya bahasa yang digunakan oleh pengguna sehingga dibutuhkan pendekatan yang lebih efisien dan sistematis dalam menganalisis ulasan pengguna[8]. Peran data mining dapat diigunakan untuk menganalisis kebiasaan pelanggan, memproyeksikan penjualan, mengatur persediaan, dan mendeteksi risiko usaha. Penerapan data mining dapat meningkatkan efisiensi operasional, memperkuat hubungan dengan pelanggan, dan memberikan keunggulan kompetitif yang mendukung daya saing UKM di era digital [9]. Dengan menerapkan data mining, peneliti dapat menemukan topik utama, mengamati perubahan sentimen, serta mendapatkan informasi tersembunyi atau pola yang bisa membantu dalam membuat keputusan [10]

Salah satu tujuan utama data mining adalah menemukan pola, hubungan, atau struktur dalam data yang sebelumnya tidak terlihat [11] pendekatan yang dapat digunakan dalam menganalisis ulasan berbasis teks adalah melalui analisis sentimen dan klasterisasi, yang memungkinkan pengelompokan ulasan berdasarkan kesamaan tema dan sentimen [12].Teknik ini memungkinkan pengelompokkan ulasan pengguna berdasarkan pola kemiripan topik dan sentimen yang terkandung di dalamnya. Dengan menerapkan metode Clustering, data ulasan dapat dikelompokkan secara otomatis ke dalam beberapa klaster utama yang merepresentasikan tema-tema ulasan dominan, seperti performa fitur, kendala teknis, kemudahan penggunaan, dan lain sebagainya. Metode ini tergolong dalam teknik unsupervised learning dan efektif dalam menangani data dalam jumlah besar secara efisien [13]. Metode Clustering dengan metode K-means merupakan salah satu teknik yang terbukti efektif dalam pengelompokan data dalam skala besar. Dalam penelitian sebelumnya, metode ini telah digunakan untuk mengklasifikasikan pelanggan berdasarkan pola penggunaan listrik oleh PLN[14], mengelompokkan sentimen konsumen terhadap e-commerce di Indonesia [15] hingga

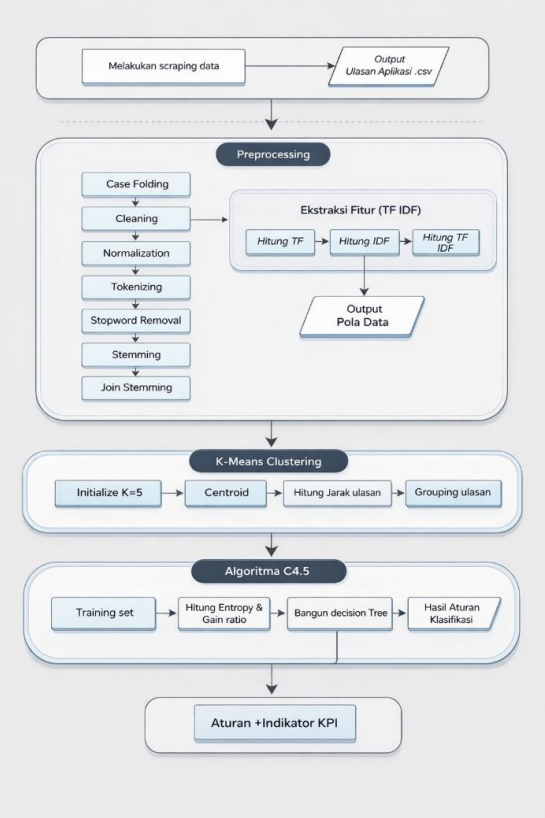
menganalisis pola kriminalitas dengan akurasi tinggi melalui evaluasi indeks Davies-Bouldin [16]. Hasil studi ini menunjukkan bahwa metode K-Means mampu memberikan segmentasi yang jelas dan informatif untuk pengambilan keputusan strategis. Setelah cluster terbentuk, algoritma C4.5 diterapkan untuk mengklasifikasikan ulasan dalam tiap cluster ke dalam kategori spesifik berdasarkan fitur yang relevan, seperti kendala teknis, kemudahan penggunaan, atau permintaan fitur tambahan [17]

Metode C4.5 digunakan sebagai teknik klasifikasi untuk mengkategorikan ulasan ke dalam kelas-kelas tertentu berdasarkan fitur yang relevan[18]. Pemilihan parameter yang benar dapat meningkatkan akurasi, mengurangi overfitting, dan mempercepat proses klasifikasi. Oleh karena itu, pengoptimalan parameter pada algoritma C4.5 menjadi langkah krusial agar hasil data mining lebih efisien, dapat diandalkan, dan sesuai dengan sifat data yang dianalisis [19]. Kombinasi metode k-means dan algoritma C4.5 terbukti efektif dalam pengambilan keputusan sesuai dengan penelitian yang mampu mengelompokkan siswa sesuai karakteristik tertentu dan mengklasifikasikan tingkat risiko mereka untuk putus sekolah. Analisis ini memberikan data berguna bagi sekolah untuk melakukan intervensi yang sesuai, seperti pembinaan akademik, konseling, atau bantuan keuangan, sehingga peluang siswa untuk melanjutkan pendidikan dapat lebih terpelihara dengan kedua metode ini [20]Hal ini menujukkan bahwa penerapan data mining tidak hanya pada dunia Pendidikan namun pada pada potensi lainnya seperti analisis berbasis data dalam konteks evaluasi layanan digital yang lebih komprehensif sehingga mampu mendukung pengambilan keputusan yang lebih terarah dan berbasis bukti [21]

Meskipun berbagai penelitian telah menerapkan metode K-Means dan algoritma C4.5 pada berbagai bidang seperti pendidikan, kelistrikan, dan e-commerce, penelitian yang mengintegrasikan kedua metode tersebut dalam analisis ulasan aplikasi internal industri otomotif masih terbatas. Sebagian besar studi sebelumnya juga cenderung menggunakan pendekatan tunggal, baik clustering maupun klasifikasi saja, sehingga belum menghasilkan analisis yang terintegrasi. Penelitian ini menawarkan kontribusi dengan mengombinasikan K-Means Clustering untuk mengidentifikasi pola dan tema dominan ulasan, serta C4.5 untuk mengklasifikasikan hasil pengelompokan ke dalam kategori yang lebih spesifik. Pendekatan ini memberikan analisis yang lebih komprehensif dan terstruktur dalam mendukung pengambilan keputusan berbasis data pada evaluasi layanan aplikasi di sektor otomotif. Berdasarkan analisa, penelitian ini bertujuan memanfaatkan ulasan pengguna aplikasi SFID sebagai sumber data untuk dianalisis secara sistematis melalui pendekatan data mining dengan metode teknik K-Means Clustering untuk mengelompokkan ulasan berdasarkan tema dan kesan pengguna, serta metode C4.5 untuk mengklasifikasikan ulasan ke dalam kategori yang relevan. Kombinasi kedua teknik ini memungkinkan analisis yang lebih lengkap, sehingga mempermudah pengambilan keputusan berbasis data, seperti menentukan prioritas pengembangan aplikasi, memperbaiki fitur, dan meningkatkan pelayanan secara tepat sasaran. sehingga diharapkan penelitian ini memberikan manfaat secara teoretis dengan menambah wawasan mengenai penerapan metode K-Means dalam pengelompokan ulasan pengguna, transformasi data teks menjadi data numerik, serta pemanfaatan metode unsupervised learning untuk menemukan pola dalam data ulasan, dan secara praktis membantu perusahaan serta pengembang aplikasi Mitsubishi SFID dalam memahami pengalaman pengguna, mengidentifikasi keluhan secara lebih cepat, mempermudah pengelompokan permasalahan yang sering muncul, serta menjadi dasar dalam peningkatan fitur dan kualitas layanan aplikasi sesuai dengan kebutuhan pengguna. Hasil penelitian ini dapat dimanfaatkan oleh Mitsubishi untuk mengklasifikasikan ulasan pengguna ke dalam empat cluster utama, yaitu Urgent, Medium, Low, dan Tidak Relevan, sehingga perusahaan dapat menentukan prioritas penanganan dan pengembangan aplikasi secara lebih terarah dan berbasis data.

2. Metode Penelitian

Proses klasifikasi dilakukan dengan menggunakan dua pendekatan utama, yaitu metode K-Means Clustering untuk mengelompokkan data ulasan berdasarkan kesamaan karakteristik teks, serta algoritma C4.5, dimulai dengan tahap scrapping data dan kemudian dilanjutkan dengan tahap preproseccing serta diolah dengan metode K-means Clustering dan Algoritma C4.5. Adapun tahapan proses penelitian ini dapat dilihat pada gambar 1:



Gambar 1. Kerangka Penelitian

Berdasarkan kerangka penelitian yang telah ditetapkan, penelitian ini dilaksanakan secara terstruktur melalui tahapan-tahapan yang saling berkaitan, dimulai dari proses pengumpulan data hingga menghasilkan output sesuai dengan tujuan penelitian. Dataset yang digunakan berupa ulasan pengguna aplikasi Mitsubishi SFID yang diperoleh dari platform Google Play Store dengan jumlah 300 data. Data dikumpulkan menggunakan teknik scraping, kemudian disimpan dalam format spreadsheet untuk mempermudah proses pengolahan dan analisis lebih lanjut.

Tahap awal penelitian adalah preprocessing data yang bertujuan untuk meningkatkan kualitas data agar lebih rapi, bersih, dan siap digunakan pada tahap analisis. Proses ini diawali dengan case folding, yaitu menyeragamkan seluruh teks menjadi huruf kecil. Selanjutnya dilakukan cleaning untuk menghapus elemen yang tidak relevan, seperti data duplikat, ulasan kosong, emotikon, angka, tanda baca, URL, serta spasi berlebih. Tahap berikutnya adalah tokenizing, yaitu memecah teks menjadi unit kata. Kemudian dilakukan stopword removal untuk menghilangkan kata-kata umum yang tidak memiliki makna penting, serta stemming untuk mengubah kata berimbuhan menjadi bentuk kata dasar sesuai dengan kaidah bahasa Indonesia. Proses preprocessing ditutup dengan join stemming, yaitu menggabungkan kembali kata-kata hasil stemming sebagai representasi akhir teks yang akan dianalisis.

Seluruh tahapan pengolahan data dilakukan menggunakan bahasa pemrograman Python dengan dukungan beberapa pustaka seperti Pandas, NumPy, dan Scikit-learn. Dengan rangkaian proses tersebut, data yang dihasilkan menjadi lebih terstruktur dan siap digunakan dalam penerapan metode K-Means Clustering dan algoritma C4.5 untuk memperoleh hasil analisis yang akurat dan dapat dipertanggungjawabkan

TF IDF

Hasil pemprosesan akan dilakukan ektraksi fitur dimana hasil data preprocessing akan di representasikan ke numeric supaya clustering dapat mengolah data tersebut karena algoritma tidak dapat menganalisasi secara texs langsung, proses ini merupakan salah satu cara untuk melakukan penghitungan bobot setiap kata yang umum digunakan dalam suatu

dokumen teks sehingga dapat mengetahui seberapa penting sebuah kata untuk di olah. Document frequency digunakan untuk mengukur seberapa banyak dokumen yang mengandung suatu term, sehingga dapat menunjukkan kemunculan atau distribusi suatu term dalam corpus. Adapun persamaan df adalah sebagai berikut:

$DF_{(t)}$ = Jumlah yang banyak mengandung data t .....(1)

DF merujuk pada jumlah dokumen yang banyak mengandung kata tertentu (t) dalam suatu kumpulan data Sedangkan untuk persamaan Term Frequency (TF) adalah sebagai berikut, di mana TF menyatakan frekuensi kemunculan suatu term pada dokumen tertentu yang digunakan sebagai dasar dalam proses pembobotan teks.

$$TF_{t,d} = \frac{Jumlah\ kemunculan\ kata\ t\ di\ dokumen\ d}{Total\ Jumlah\ kata\ di\ dokumen\ d} \quad .....(2)$$

Term Frequency (TF) merupakan ukuran yang digunakan untuk menunjukkan seberapa sering suatu kata tertentu (ttt), misalnya “error”, muncul dalam suatu dokumen tertentu (ddd), misalnya Ulasan 1. Nilai TF dihitung dengan membandingkan jumlah kemunculan kata tersebut dalam satu dokumen terhadap total seluruh kata. Nilai TF dan IDF dikalikan untuk mendapatkan bobot kata (term weight) pada setiap kata dalam dokumen. Bobot ini menunjukkan seberapa penting suatu kata dalam dokumen dibandingkan dengan seluruh kumpulan dokumen. Persamaan TF-IDF adalah sebagai berikut:[22]

$$W_{d.t} = tf_{dt}\ x\ idf_{dt} \quad .....(3)$$

Dalam perhitungan pembobotan teks menggunakan metode TF-IDF, D menyatakan dokumen ke-d yang sedang dianalisis, sedangkan t merupakan term atau kata ke-t yang terdapat dalam dokumen tersebut. W menunjukkan bobot suatu dokumen (d) terhadap term (t), yang diperoleh dari hasil perkalian antara nilai tf dan idf. Nilai tf (term frequency) adalah jumlah kemunculan suatu term dalam dokumen tertentu, yang menggambarkan tingkat kepentingan kata tersebut di dalam dokumen. Sementara itu, idf (Inverse Document Frequency) digunakan untuk mengukur tingkat keunikan suatu term dalam keseluruhan kumpulan dokumen. Nilai idf dihitung berdasarkan df (document frequency), yaitu jumlah dokumen yang mengandung term tersebut. Semakin sedikit dokumen yang memuat suatu term (nilai df kecil), maka nilai idf akan semakin besar, sehingga bobot akhir (W) menjadi lebih tinggi. Dengan demikian, pembobotan TF-IDF mampu menyeimbangkan frekuensi kemunculan kata dalam dokumen dengan tingkat penyebarannya dalam seluruh koleksi dokumen

K-Means Clustering

Dalam penelitian ini, hasil matriks skor TF-IDF akan dilanjutkan ke proses clustering K-Means dimana K-Means mengelompokan ulasan berdasarkan kedekatan nilai skor TF-IDF antar ulasan, proses clustering K-Means mengelompokkan ulasan yang memiliki pola kata dan bobot fitur dalam satu kelompok atau cluster yang sama sehingga ulasan dengan kemiripan karakteristik dapat dianalisa secara bersamaan. Proses awal dimulai dengan ditentukannya jumlah cluster dimana pada penelitian ini ditentukan jumlah cluster sebanyak 4 (K=4) yang merepresentasikan tingkat prioritas dalam pengembangan aplikasi Mitsubishi SFID. Setelah jumlah klaster optimal ditentukan, langkah selanjutnya dalam penerapan algoritma K-Means adalah menghitung jarak antara setiap titik data dengan pusat klaster (centroid). Pada tahap perhitungan jarak ini menggunakan rumus dengan Euclidean Distance sebagai berikut[23]

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad .....(4)$$

Dalam perhitungan metode K-Means Clustering, d(x,y) menyatakan jarak antara data pada titik x dan titik y. Nilai jarak ini digunakan untuk menentukan kedekatan suatu data terhadap pusat cluster. x merepresentasikan titik data objek yang akan dikelompokkan,

sedangkan y merupakan titik data centroid atau pusat cluster yang menjadi acuan pengelompokan. Sementara itu, I menunjukkan jumlah atribut atau variabel yang dimiliki setiap data. Perhitungan jarak umumnya menggunakan rumus Euclidean Distance, yaitu dengan menjumlahkan selisih kuadrat setiap atribut antara titik x dan y, kemudian diakarkan. Semakin kecil nilai d(x,y), maka semakin dekat data tersebut dengan centroid dan semakin besar kemungkinan data tersebut menjadi anggota cluster tersebut.

Adapun untuk menghitung centroid baru, dilakukan dengan menggunakan persamaan tertentu yang bertujuan untuk menentukan titik pusat klaster berdasarkan rata-rata nilai seluruh data yang tergabung dalam klaster tersebut. Perhitungan centroid baru ini dilakukan setelah proses pengelompokan data pada iterasi sebelumnya. Persamaan yang digunakan untuk menghitung centroid baru disajikan sebagai berikut:

$$C_i^{new} = (\frac{1}{N}\sum_{j=1}^{N_1} x_j^{(1)} , \frac{1}{N_1}\sum_{j=1}^{N_1} x_j^2 , \dots, \frac{1}{N}\sum_{j=1}^{N_1} x_j^{(n)} \ ) \qquad \text{.....(5)}$$

Dalam proses iterasi metode K-Means Clustering, centroid akan diperbarui pada setiap tahap perhitungan. Simbol tersebut menyatakan centroid baru untuk klaster ke-i, yaitu titik pusat terbaru yang diperoleh setelah menghitung rata-rata seluruh data yang termasuk dalam klaster tersebut. Nilai centroid baru dihitung berdasarkan rata-rata setiap dimensi atau atribut dari seluruh anggota klaster. Sementara itu, nilai data pada dimensi pertama, kedua, dan seterusnya untuk data ke-j dalam klaster i merepresentasikan nilai atribut dari masing-masing data yang menjadi anggota klaster tersebut. Dengan demikian, centroid baru diperoleh dengan menjumlahkan seluruh nilai atribut pada setiap dimensi dalam klaster i, kemudian dibagi dengan jumlah data yang ada dalam klaster tersebut. Proses ini dilakukan secara berulang hingga posisi centroid stabil dan tidak lagi mengalami perubahan signifikan.

Algoritma C4.5

Algoritma data mining C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan dan bersifat prediktif. Langkah awal dari algoritma C4.5 adalah menghitung nilai entropi. Tahap pertama dilakukan dengan menentukan nilai entropi total pada suatu kasus. Perhitungan entropi dilakukan menggunakan rumus sebagai berikut[17]:

$$Entropy\ (S) = \sum_{i=1}^{n} -pi \ * log_2 \ \ pi \qquad \text{.....(6)}$$

Dalam perhitungan entropy pada algoritma klasifikasi seperti C4.5, S menyatakan himpunan kasus atau seluruh data yang sedang dianalisis. n menunjukkan jumlah partisi dari himpunan S, yaitu jumlah kelas atau kategori yang terbentuk dalam data tersebut. Sementara itu, p_i merupakan proporsi dari partisi ke-i (S_i) terhadap keseluruhan himpunan S, yang dihitung berdasarkan perbandingan jumlah data pada kelas ke-i dengan total seluruh data. Nilai proporsi ini digunakan dalam perhitungan entropy untuk mengukur tingkat ketidakpastian atau keberagaman data dalam suatu himpunan. Semakin merata distribusi proporsi antar kelas, maka nilai entropy akan semakin tinggi, yang menunjukkan tingkat ketidakpastian yang lebih besar.

Di mana S merupakan himpunan kasus, A adalah atribut, n adalah jumlah partisi, dan p_i adalah proporsi jumlah kasus pada partisi ke-i terhadap keseluruhan kasus dalam S. S merupakan sekumpulan data kasus yang akan dianalisis, A adalah atribut yang digunakan dalam setiap kasus, n menyatakan jumlah partisi dari atribut A, dan p_i menunjukkan proporsi data pada masing-masing partisi. Setelah nilai entropi diperoleh, langkah selanjutnya adalah menghitung nilai information gain untuk menentukan akar (root) pada pohon keputusan. Perhitungan nilai gain dilakukan dengan menggunakan rumus berikut:[17]

$$Gain(S,A) = Entropy(S) - \sum_{i-1}^{n} \frac{|Si|}{|S|} * Entropy\,(Si) \quad \text{.....(7)}$$

Dalam proses perhitungan information gain pada algoritma C4.5, S merepresentasikan keseluruhan kumpulan data atau himpunan kasus yang sedang dianalisis. A adalah atribut yang digunakan untuk membagi data tersebut menjadi beberapa bagian. Jumlah pembagian yang dihasilkan dari atribut A dinyatakan dengan n, yaitu banyaknya partisi yang terbentuk setelah proses pemisahan dilakukan. Simbol $|S_i|$ menunjukkan jumlah data yang terdapat pada partisi ke-i sebagai hasil pembagian berdasarkan atribut A, sedangkan |S| menyatakan total seluruh data dalam himpunan S sebelum dilakukan pemisahan. Nilai-nilai ini digunakan untuk menghitung tingkat pengurangan ketidakpastian (entropy) sehingga dapat menentukan atribut yang paling optimal dalam membentuk struktur pohon keputusan.

3. Hasil dan Pembahasan

Tahap ini mempresentasikan hasil pengolahan dan analisis data ulasan pengguna aplikasi yang telah menjalani proses prapemrosesan serta transformasi menjadi format numeric. Analisis dilaksanakan dengan menggunakan metode pengelompokan dan pengklasifikasian untuk mengategorikan ulasan berdasarkan aspek aplikasi, layanan, dan sentimen. Selanjutnya, hasil yang didapat dibahas untuk menemukan pola penilaian pengguna dan menentukan kategori prioritas yang bisa dijadikan sebagai dasar untuk evaluasi dan perbaikan aplikasi.

Preprocessing

Hasil pengambilan data ulasan pengguna dari *Google Play Store* menunjukkan bahwa data yang diperoleh masih berupa data mentah yang mengandung berbagai elemen tidak relevan, seperti simbol, emoji, duplikasi teks, dan data kosong. Untuk mengatasi hal tersebut, dilakukan tahapan preprocessing pada kolom content agar kualitas data meningkat sebelum memasuki tahap analisis. Proses preprocessing yang dilakukan mencakup penyeragaman huruf, pembersihan teks, pemecahan kata, penghapusan kata umum, serta pengubahan kata berimbuhan ke bentuk dasarnya, sehingga menghasilkan teks yang lebih rapi dan konsisten.

Berdasarkan hasil *preprocessing*, teks ulasan yang sebelumnya tidak terstruktur berhasil ditransformasikan menjadi data teks yang lebih informatif dan representatif terhadap opini pengguna. Pengurangan kata-kata yang tidak bermakna serta penyatuan bentuk kata yang memiliki makna serupa terbukti mampu meminimalkan gangguan (*noise*) dalam data. Data hasil preprocessing ini selanjutnya digunakan sebagai dasar dalam proses ekstraksi fitur dan analisis lanjutan menggunakan metode clustering dan klasifikasi untuk menggali pola penilaian pengguna terhadap aplikasi. Adapun Data bersih hasil *preprocessing* adalah sebagai berikut:

Tabel 1. Hasil *Preprocessing* setelah dilakukan *stemming*

Data	Hasil *Preprocessing*
D1	aplikasi berat masuk e learning bikin simpel aja pake captcha captcha an ribet tau pikir nya keren kali ya baik bos
D2	tolong baik simulasi kredit simulasi kredit fungsi hitung simulasi kredit hasil error butuh dealing konsumen tolong banget ya bantu cmo dipo star finansial lambat respon cmo dipo lambat kes bantu sales
D3	guna sales krna sfid tau product knowledge simpan data base
D4	aplikasi ulas bintang rasa aplikasi logout fungsi kamera kadang ambil data galeri kadang tunjuk betapa buruk aplikasi mesti google cabut aplikasi aplikasi
D5	lot hang suka log out pokok cacat nih aplikasi habis baru buka bagus parah pake aplikasi tolol susah aja
D6	mohon ktb aplikasi berat hp support dana ganti hp haha login aja pinjam hp keluarga
D7	tolong ya d baca ulas nya bintang gue kalo sok an digital dech bikin pusing bentar baru bentar baru ngerepotin banget
D8	login hapus instal ulang ketuk tahan aplikasi pilih info detail aplikasi pilih simpan cache pilih hapus simpan data log in
D9	aplikasi susah sales banyak uninstall aduh vendor aplikasi benefit ecek dikit uninstall
D10	aplikasi receh hari buka besok langsung error uninstal instal baik aplikasi
D11	aplikasi sampah buka error suruh sih cache suruh ganti password suruh uninstal instal ganti hp suport heh bbm naikin gaji dlu suruh ganti hp sambo kau
D12	bantu depan mantap utk program dukung layan puas langgan customer happy mitsubishi emang keren
D13	lot instal ulang ulang ajah dahh

Data	Hasil *Preprocessing*
D14	bintang klau bener ksh bintang
D15	cape banget jual aja bikin ribet sih aduh sumpah aplikasi bobot matching lead aja susah banget udah suka log out sumpah bikin orang susah era digital harus nya aplikasi manis mudah kerja nya bikin sulit mohon ajar mmksi thanks
D16	tugas beban sales gun administrasi sales input sales job desk nya jual administrator jual aplikasi tu jual otomotif mmksi ubah tunjang jual langsung tangan jual kena saring baca ni mmksi
D17	gin deh titip sen aja kembang aplikasi sampein orang mitsubishi pusat aplikasi nambah kerja sales fight lapang dapat beban input data fungsi admin kantor admin gaji dealer ngasih gaji sales percaya coba aja muter daerah deh dah gitu aja kali aja baca orang mitsubishi
D18	cape aplikasi log out sendri kalo masuk hapus aplikasi download masuk bantu mudah kerja serba sulit tolong baik masuk baca aja koreksi jalan terima kasih
D19	aplikasi bener bagus min hapenya canggih bagus aplikasi aja super lot udah gitu logout susah login susah sales sinkron warranty bareng konsumen mmid udah gitu pdi expired gimana kerja cepet aplikasi aja lot payah bener tim itnya kaya anak ajar it aplikasi super lot kasi konsumen warranty aktif service
D20	yth admin pkt pdi kerja sales lapang sulit cari order jaman pandemi kerja kaya gin kerja divisi sales tolong benah kerja sales tambah tambahin sedang kerja divisi rang

Pada Tabel 1 ditampilkan data teks yang telah melalui proses preprocessing dan berada dalam kondisi bersih. Data tersebut selanjutnya digunakan sebagai input dalam tahap pembobotan kata menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Proses pembobotan TF-IDF dilakukan terhadap teks hasil preprocessing dengan menyesuaikan tiga variabel aspek yang digunakan dalam penelitian, yaitu aspek aplikasi, aspek layanan, dan aspek sentimen. Setiap ulasan pengguna direpresentasikan dalam bentuk vektor numerik TF-IDF, kemudian nilai bobot kata dikelompokkan sesuai dengan aspek yang relevan. Kata-kata yang berkaitan dengan fitur, performa, dan kemudahan penggunaan aplikasi dipetakan ke aspek aplikasi, sementara kata-kata yang mencerminkan pelayanan, respons, dan interaksi pengguna dipetakan ke aspek layanan, kata-kata yang menunjukkan ekspresi emosi dan penilaian pengguna dipetakan ke aspek sentimen. Adapun hasil pola analisi data sampel terhadap variabel adalah sebagai berikut:

Tabel 2. Hasil Pola Analisis data sampel terhadap variabel

Data	Aspek Aplikasi	Aspek Layanan	Aspek Sentimen
D1	1.0782	0	0
D2	0.3046	1.0694	1.0694
D3	0.1204	0	0
D4	0.3905	0	0
D5	0.7952	0	0
D6	0.1772	0.2	0.2
D7	0.4	0.1398	0.1398
D8	0.917	0	0
D9	0.4044	0.2602	0.2602
D10	0.684	0	0
D11	1.1068	0.2	0.2
D12	0.2602	0.425	0.425
D13	0.5592	0	0
D14	0	0	0
D15	0.335	0.2694	0.2694
D16	0.3578	0.5388	0.5388
D17	0.1952	1.7444	1.7444
D18	0.418	0.4092	0.4092
D19	0.9497	0.5648	0.5648
D20	0.2	0.6183	0.6183

Berdasarkan hasil pembobotan TF-IDF pada Tabel 1, terlihat bahwa tiga aspek yang dianalisis, yaitu aspek aplikasi, aspek layanan, dan aspek sentimen, menunjukkan perbedaan pola dominasi pada setiap data ulasan. Perbedaan ini mencerminkan variasi fokus dan penekanan pengguna dalam menyampaikan pendapatnya, baik terkait fungsi dan performa aplikasi, kualitas layanan yang diterima, maupun ekspresi sentimen atau penilaian yang diberikan. Hal tersebut menunjukkan bahwa setiap ulasan memiliki karakteristik yang berbeda-beda sesuai dengan pengalaman dan persepsi masing-masing pengguna. Beberapa ulasan menunjukkan nilai TF-IDF yang lebih tinggi pada aspek aplikasi, yang mengindikasikan fokus pengguna terhadap fungsi dan kualitas aplikasi, sementara nilai pada aspek layanan dan sentimen relatif rendah atau nol. Sebaliknya, terdapat ulasan dengan bobot yang lebih besar pada aspek layanan dan sentimen, yang mencerminkan pengalaman pengguna terhadap layanan serta ekspresi opini yang lebih kuat. Selain itu, beberapa data memiliki nilai yang relatif

seimbang pada ketiga aspek, menunjukkan bahwa ulasan tersebut membahas aplikasi, layanan, dan sentimen secara bersamaan. Variasi pola ini menunjukkan bahwa representasi TF-IDF berbasis tiga aspek mampu menangkap karakteristik ulasan pengguna secara komprehensif dan dapat digunakan sebagai dasar analisis lanjutan.

K-Means

Representasi numerik ulasan pengguna yang diperoleh dari pembobotan TF-IDF digunakan sebagai input dalam proses clustering menggunakan algoritma K-Means untuk mengidentifikasi pola kesamaan data. Pada proses pengelompokan menggunakan metode K-Means, satu kali iterasi awal dilakukan untuk menganalisis pola pengelompokan data berdasarkan nilai dari aspek aplikasi, aspek layanan, dan aspek sentimen. Hasil dari iterasi pertama menunjukkan bahwa ulasan mulai terdistribusi dalam kelompok berdasarkan kedekatan nilai di ketiga aspek ini. Ulasan yang memiliki nilai TF-IDF yang lebih signifikan pada aspek aplikasi cenderung berada dalam satu kelompok yang sama, sementara ulasan dengan skor lebih tinggi pada aspek layanan dan sentimen membentuk kelompok yang berbeda. Di samping itu, terdapat beberapa data dengan nilai yang cukup seimbang di ketiga aspek yang berada di antara kedua kelompok tersebut. Temuan ini menandakan bahwa meskipun hanya melalui satu iterasi, metode K-Means sudah bisa menangkap pola awal dari kesamaan karakteristik ulasan pengguna, yang kemudian akan disempurnakan pada iterasi selanjutnya hingga mencapai kondisi stabil. Adapun hasil jarak pada iterasi dan menghasilkan nilai yng konvergen adalah sebagai berikut:

Tabel 3. Hasil K-means

Data	C1	C2	C3	C4	Kelas
D1	2.155184	0.809018	0.226428	0.880219	3
D2	0.480421	0.961975	1.565797	1.144872	1
D3	1.993867	1.151406	0.737841	0.400327	4
D4	1.994619	0.975040	0.468610	0.384461	4
D5	2.063029	0.813553	0.077515	0.634228	3
D6	1.708362	0.928932	0.719250	0.132200	4
D7	1.798225	0.814506	0.480911	0.209019	4
D8	2.098513	0.799417	0.076513	0.736850	3
D9	1.629022	0.694920	0.554565	0.125900	4
D10	2.036462	0.841781	0.179050	0.547445	3
D11	1.909841	0.539294	0.343641	0.832644	3
D12	1.388654	0.717285	0.814038	0.233972	4
D13	2.013554	0.889094	0.301245	0.462665	3
D14	2.005289	1.240938	0.858029	0.461333	4
D15	1.610917	0.743222	0.619403	0.058022	4
D16	1.232411	0.593041	0.871738	0.402090	4
D17	0.480421	1.830897	2.508610	2.100821	1
D18	1.420939	0.575437	0.689230	0.252871	4
D19	1.381298	0.000000	0.757335	0.797645	2
D20	1.116365	0.753508	1.056253	0.512668	4

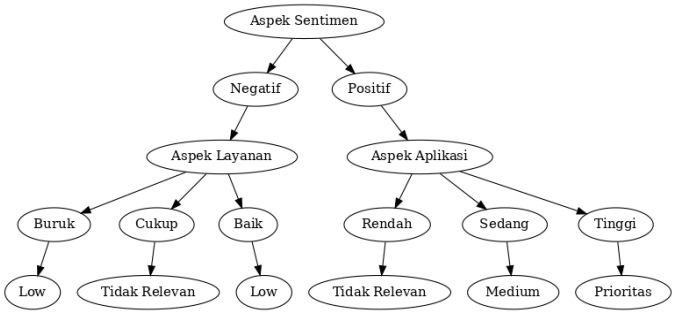
Pada tabel 3 ditampilkan hasil pengelompokan 20 data ulasan menggunakan algoritma K-Means yang menghasilkan empat cluster, yaitu C1 (Prioritas), C2 (Medium), C3 (Low), dan C4 (Tidak Relevan). Hasil clustering menunjukkan bahwa sebagian besar data, yaitu 11 ulasan, tergabung dalam cluster C4 (Tidak Relevan). Hal ini mengindikasikan bahwa mayoritas ulasan memiliki tingkat relevansi yang rendah terhadap aspek aplikasi, layanan, dan sentimen yang dianalisis. Selanjutnya, 6 data termasuk dalam cluster C3 (Low), yang mengindikasikan ulasan dengan tingkat kepentingan rendah. Sementara itu, cluster C1 (Prioritas) terdiri dari 2 data dan cluster C2 (Medium) hanya terdiri dari 1 data, yang menandakan bahwa ulasan dengan tingkat urgensi tinggi dan sedang relatif sedikit. Pola ini menunjukkan bahwa algoritma K-Means mampu mengelompokkan data ulasan berdasarkan tingkat kemiripan karakteristiknya secara jelas, sehingga hasil clustering dapat digunakan sebagai dasar analisis lanjutan pada tahap klasifikasi.

Algoritma C4.5

Hasil clustering menggunakan algoritma K-Means selanjutnya digunakan sebagai dasar dalam proses klasifikasi menggunakan algoritma C4.5. Sebelum dilakukan perhitungan C4.5,

setiap variabel numerik pada aspek aplikasi, aspek layanan, dan aspek sentimen terlebih dahulu dikategorikan menggunakan metode Sturges untuk mengubah data kontinu menjadi data diskrit. Proses kategorisasi ini menghasilkan interval nilai yang merepresentasikan tingkat kepentingan tertentu, sehingga data lebih sesuai untuk proses pembentukan pohon keputusan. Setelah tahap kategorisasi dilakukan, data hasil K-Means tersebut kemudian dihitung menggunakan algoritma C4.5 untuk menentukan aturan klasifikasi dan menghasilkan pohon keputusan yang menggambarkan hubungan antar aspek dalam menentukan kelas prioritas ulasan pengguna.

Berdasarkan hasil perhitungan algoritma C4.5 terhadap 20 data ulasan, diperoleh nilai entropi awal sebesar 1,5438 yang menunjukkan tingkat ketidakpastian kelas sebelum dilakukan pemisahan atribut. Perhitungan information gain menunjukkan bahwa Aspek Aplikasi memiliki nilai gain tertinggi sebesar 0,7439, diikuti oleh Aspek Layanan sebesar 0,6559, dan Aspek Sentimen sebesar 0,3110. Hasil ini mengindikasikan bahwa Aspek Aplikasi merupakan atribut paling berpengaruh dalam menentukan kelas prioritas ulasan pengguna dan dipilih sebagai node akar (root) pada pohon keputusan C4.5. Pembagian data berdasarkan kategori Aspek Aplikasi (rendah, sedang, dan tinggi) mampu mengurangi ketidakpastian kelas secara signifikan dibandingkan atribut lainnya. Dengan demikian, hasil C4.5 menunjukkan bahwa karakteristik ulasan pengguna lebih dominan ditentukan oleh pembahasan terkait aplikasi dibandingkan aspek layanan dan sentimen dalam proses penentuan tingkat prioritas ulasan. Pohon keputusan yang terbentuk dapat dirangkum ke dalam aturan berikut:



Gambar 1. Hasil Pohon keputuasan

Berdasarkan hasil pembobotan TF-IDF pada Tabel 1, terlihat bahwa tiga aspek yang dianalisis, yaitu aspek aplikasi, aspek layanan, dan aspek sentimen, menunjukkan pola dominasi yang berbeda pada setiap data ulasan. Beberapa ulasan memiliki nilai TF-IDF yang lebih tinggi pada aspek aplikasi, yang mengindikasikan bahwa pengguna lebih menyoroti fungsi, performa, dan kualitas sistem aplikasi. Pada ulasan tersebut, nilai aspek layanan dan sentimen cenderung rendah atau mendekati nol. Sebaliknya, terdapat ulasan dengan bobot yang lebih dominan pada aspek layanan dan sentimen, yang mencerminkan fokus pengguna terhadap kualitas pelayanan, respons, interaksi, serta ekspresi opini atau emosi terhadap aplikasi. Selain itu, ditemukan pula ulasan dengan distribusi nilai yang relatif seimbang pada ketiga aspek, yang menunjukkan bahwa isi ulasan membahas aplikasi, layanan, dan sentimen secara bersamaan. Variasi pola pembobotan ini menunjukkan bahwa representasi TF-IDF berbasis tiga aspek mampu mengidentifikasi karakteristik utama dalam setiap ulasan secara lebih komprehensif. Dengan demikian, hasil pembobotan tersebut dapat dijadikan sebagai dasar yang kuat untuk tahap analisis lanjutan, seperti proses clustering dan klasifikasi.

Setiap aturan disusun dalam bentuk IF–THEN, di mana bagian IF menunjukkan kondisi atribut yang terpenuhi, sedangkan bagian THEN menunjukkan kelas hasil keputusan. Atribut yang digunakan dalam pembentukan aturan meliputi Aspek Sentimen, Aspek Layanan, dan Aspek Aplikasi, yang masing-masing berperan sebagai dasar penentuan tingkat prioritas.

IF Aspek Aplikasi = Rendah AND Aspek Layanan = Baik THEN Prioritas
IF Aspek Aplikasi = Rendah AND Aspek Layanan ≠ Baik THEN Tidak Relevan
IF Aspek Aplikasi = Sedang AND Aspek Sentimen = Positif THEN Low
IF Aspek Aplikasi = Sedang AND Aspek Sentimen = Negatif THEN Tidak Relevan
IF Aspek Aplikasi = Tinggi AND Aspek Sentimen = Positif THEN Medium
IF Aspek Aplikasi = Tinggi AND Aspek Sentimen = Negatif THEN Low

Hasil Evaluasi

Setelah proses pengelompokan data menggunakan algoritma K-Means dan dilanjutkan dengan klasifikasi menggunakan algoritma C4.5, diperoleh model klasifikasi yang merepresentasikan hubungan antara aspek aplikasi, aspek layanan, dan aspek sentimen dalam menentukan tingkat prioritas ulasan pengguna. Model yang terbentuk kemudian diuji untuk menilai kemampuan prediksinya terhadap data ulasan. Evaluasi kinerja dilakukan untuk mengukur tingkat keakuratan model dalam mengklasifikasikan data.

```
=== HASIL ALGORITMA C4.5 (Decision Tree) ===
Akurasi: 0.8889

Laporan Klasifikasi:
              precision    recall  f1-score   support

         Low       0.83      0.83      0.83         6
      Medium       0.95      0.95      0.95        21
Tidak Relevan      0.60      0.60      0.60         5
      Urgent       1.00      1.00      1.00         4

    accuracy                           0.89        36
   macro avg       0.85      0.85      0.85        36
weighted avg       0.89      0.89      0.89        36
```

Gambar 2. Hasil akurasi pada proses klasifikasi

Hasil evaluasi menunjukkan bahwa model klasifikasi C4.5 memperoleh tingkat akurasi sebesar 88,89%, yang termasuk dalam kategori cukup baik. Nilai akurasi tersebut mengindikasikan bahwa model mampu mengklasifikasikan ulasan pengguna secara tepat berdasarkan pola kata yang dihasilkan dari pembobotan TF-IDF serta hasil pengelompokan menggunakan K-Means. Tingginya tingkat akurasi ini menunjukkan bahwa kombinasi representasi fitur berbasis tiga aspek dan pendekatan klasifikasi yang digunakan efektif dalam menangkap karakteristik ulasan pengguna. Dengan demikian, model C4.5 yang diusulkan dapat digunakan sebagai alat pendukung dalam menentukan tingkat prioritas ulasan pengguna secara andal

4. Kesimpulan

Berdasarkan hasil penelitian, klasifikasi Ulasan Pengguna Aplikasi Mitsubishi Menggunakan Pendekatan K-Means Clustering dan Algoritma C4.5 mampu menunjukkan bahwa kombinasi tersebut efektif dalam mengklasifikasikan ulasan pengguna ke dalam kategori prioritas yang relevan, dibuktikan dengan hasil evaluasi menunjukkan tingkat akurasi sebesar 88,89%. Selain itu, hasil penelitian ini dapat dijadikan sebagai solusi alternatif berbasis data bagi pihak pengembang aplikasi dalam mengevaluasi kualitas layanan dan fitur aplikasi dan bagi peneliti selanjutnya disarankan untuk mengembangkan penelitian ini dengan menambahkan jumlah dan variasi data ulasan, memperkaya atribut yang digunakan, serta menerapkan teknik pemprosesan atau metode klasifikasi lain guna memperoleh hasil yang lebih optimal dan meningkatkan performa sistem klasifikasi.

Referensi

[1] Mohamad Rafi, Rama Iqbal Yudhistira Sujana, Heykel Revelyn Fahrezy, and Mohamad Zein Saleh, “Impor Mobil Suzuki dalam Perkembangan Industri Otomotif di Indonesia,” *Lokawati J. Penelit. Manaj. dan Inov. Ris.*, vol. 2, no. 6, pp. 284–295, 2024, doi: 10.61132/lokawati.v2i6.1376.

[2] G. Kawengian, J. A. F. Kalangi, and ..., “Pengaruh Harga dan Kualitas Produk terhadap Keputusan Pembelian Mobil Xpander pada PT. Mitsubishi di Dealer Beta Berlian Winangun Manado,” *Productivity*, vol. 3, no. 6, pp. 531–535, 2022, [Online]. Available: https://ejournal.unsrat.ac.id/index.php/productivity/article/view/44632%0Ahttps://ejournal.unsrat.ac.id/index.php/productivity/article/download/44632/38876

[3] Mitsubishi Motors Indonesia, “Laporan Tahunan,” 2024, *Mitsubishi Motors Indonesia, Jakarta*.

[4] B. Support, “No Title,” *User Guid. SFID*, 2023.

[5] Desmond Lawrence Cook, *Program Evaluation and Review Technique: Applications in Education*, 1st ed. Routledge, 2006. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=YNscFxuJ1dYC&oi=fnd&pg=PR3&dq=application+review+as+an+evaluation

[6] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, “Analysing app reviews for software

engineering: a systematic literature review," *Empir. Softw. Eng.*, vol. 27, no. 2, 2022, doi: 10.1007/s10664-021-10065-7.

[7] F. Alqahtani and R. Orji, "Insights from user reviews to improve mental health apps," *Health Informatics J.*, vol. 26, no. 3, pp. 2042–2066, 2020, doi: 10.1177/1460458219896492.

[8] F. Broder, "Big Data Challenge: Why Manual Review Tracking Isn't Enough," Revuze Blog. Accessed: Sep. 12, 2025. [Online]. Available: https://www.revuze.it/blog/overcome-the-big-data-challenge-why-manual-review-tracking-doesnt-cut-it/

[9] S. V. Tsiu, M. Ngobeni, L. Mathabela, and B. Thango, "Applications and Competitive Advantages of Data Mining and Business Intelligence in SMEs Performance: A Systematic Review," *Businesses*, vol. 5, no. 2, p. 22, 2025, doi: 10.3390/businesses5020022.

[10] A. Dahiya, N. Gautam, and P. K. Gautam, "Data mining methods and techniques for online customer review analysis: A literature review," *J. Syst. Manag. Sci.*, vol. 11, no. 3, pp. 1–26, 2021, doi: 10.33168/JSMS.2021.0301.

[11] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[12] A. Petukhova, J. P. Matos-Carvalho, and N. Fachada, "Text clustering with large language model embeddings," *Int. J. Cogn. Comput. Eng.*, vol. 6, no. March 2024, pp. 100–108, 2025, doi: 10.1016/j.ijcce.2024.11.004.

[13] A. Idrus, N. Tarihoran, U. Supriatna, A. Tohir, S. Suwarni, and R. Rahim, "Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm," *TEM J.*, vol. 11, no. 4, pp. 1871–1876, 2022, doi: 10.18421/TEM114-55.

[14] H. W. Adji and M. S. Hakim, "Clustering Marketing Strategy Based on Applications for Power Upgrading Customers," vol. 13, no. 1, pp. 57–64, 2025, doi: 10.37641/jimkes.v13i1.3029.

[15] A. E. Widjaja, A. Fransisko, C. A. Haryani, and Hery, "Text Mining Application with K-Means Clustering to Identify Sentiments and Popular Topics: a Case Study of the three Largest Online Marketplaces in Indonesia," *J. Appl. Data Sci.*, vol. 4, no. 4, pp. 441–453, 2023, doi: 10.47738/jads.v4i4.134.

[16] R. H. Maharrani, P. D. Abda'u, and M. N. Faiz, "Clustering method for criminal crime acts using K-means and principal component analysis," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 34, no. 1, pp. 224–232, 2024, doi: 10.11591/ijeecs.v34.i1.pp224-232.

[17] E. V. Astuti, A. Afandi, and D. M. Effendi, "Classification and Clustering of Internet Quota Sales Data Using C4.5 Algorithm and K-Means," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 9, no. 2, pp. 268–283, 2023, doi: 10.26555/jiteki.v9i2.25970.

[18] Y. Asri, D. Kuswardani, W. N. Suliyanti, and C. M. Tambunan, *ALGORITMA C4.5 : Klasifikasi Titik dan Jenis Gangguan pada Jaringan Distribusi Penyulang*. 2023. [Online]. Available: https://www.google.co.id/books/edition/ALGORITMA_C4_5_KLASIFIKASI_TITIK_DAN_JEN/5FzrEAAAQBAJ?hl=en&gbpv=0&kptab=overview

[19] Y. Zhang, Y. Xin, and Q. Li, "Research on parameter selection and optimization of C4.5 algorithm based on algorithm applicability knowledge base," *Sci. Rep.*, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-11901-2.

[20] S. Anwar, "Application of K-Means and C4.5 Algorithms for dropout risk prediction in vocational high schools," *Indones. J. Multidiscip. Sci.*, vol. 4, no. 5, pp. 428–435, 2025, doi: 10.55324/ijoms.v4i5.1102.

[21] I. G. I. Sudipa, I. G. M. Darmawiguna, I. M. Dendi, and ..., *Buku Ajar Data Mining. PT*, no. April. 2024.

[22] A. Septiani and I. Budi, "Klasifikasi Ulasan Pengguna Aplikasi: Studi Kasus Aplikasi Ipusnas Perpustakaan Nasional Republik Indonesia (PNRI)," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 7, no. 4, pp. 1110–1120, 2022, doi: 10.29100/jipi.v7i4.3216.

[23] N. S. W. Al Qorn, "Penerapan Algoritma K-Means Clustering Pada Ulasan Pengguna Merdeka Mengajar Di Play Store," p. 107, 2024.