JURNAL TEKNIK INFORMATIKA

Vol. 5 No. 1 Tahun 2025
ISSN 2809-1353
DOI : 10.58794/jekin.v5i1.995

Optimasi Algoritma C5.0 dengan Teknik Ensemble Boosting untuk Peningkatan Akurasi dalam Klasifikasi Ulasan Masyarakat Terhadap Layanan BPJS Kesehatan

Mohd Rinaldi Amartha*[1], Refni Wahyuni[2], Yuda Irawan [3]
[1,2,3]Universitas Hang Tuah Pekanbaru
e-mail: *[1]amartharc@gmail.com, [2]refniabid@gmail.com, [3]yudairawan89@gmail.com

Abstract *– This research aims to improve the accuracy of sentiment classification of community reviews of BPJS Health services by optimizing the C5.0 algorithm using SMOTE and XGBoost techniques. Tests were conducted with several combinations, including C5.0, C5.0 with XGBoost, C5.0 with SMOTE, and a combination of the three. The results show that the basic C5.0 algorithm achieved an accuracy of 67.18%, the combination of C5.0 with XGBoost achieved 73.55%, C5.0 with SMOTE had an accuracy of 67.00%, while the combination of the three (C5.0, SMOTE, and XGBoost) provided the highest accuracy of 80.87%, outperforming other methods. Sentiment analysis indicates that the majority of reviews tend to be negative, highlighting consumer dissatisfaction with BPJS Kesehatan services. The significant improvement in accuracy with the application of SMOTE and XGBoost suggests that handling class imbalance and model strengthening through boosting can improve the weaknesses of the C5.0 algorithm. This clarifies the importance of ensemble strategies in complex text classification. The findings show that the use of SMOTE and XGBoost can significantly improve model performance, helping in understanding public perception more accurately.*

Keywords - C5.0 Algorithm, SMOTE, XGBoost, Ensemble, Classification

Abstrak *– Penelitian ini bertujuan untuk meningkatkan akurasi klasifikasi sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan dengan mengoptimalkan algoritma C5.0 menggunakan teknik SMOTE dan XGBoost. Pengujian dilakukan dengan beberapa kombinasi, termasuk C5.0, C5.0 dengan XGBoost, C5.0 dengan SMOTE, dan kombinasi ketiganya. Hasil menunjukkan bahwa algoritma C5.0 dasar mencapai akurasi sebesar 67.18%, kombinasi C5.0 dengan XGBoost mencapai 73.55%, C5.0 dengan SMOTE memiliki akurasi 67.00%, sementara kombinasi ketiganya (C5.0, SMOTE, dan XGBoost) memberikan akurasi tertinggi sebesar 80.87%, mengungguli metode lain. Analisis sentimen mengindikasikan bahwa mayoritas ulasan cenderung negatif, menyoroti ketidakpuasan konsumen terhadap layanan BPJS Kesehatan. Peningkatan akurasi yang signifikan dengan penerapan SMOTE dan XGBoost menunjukkan bahwa penanganan ketidakseimbangan kelas dan penguatan model melalui Boosting dapat memperbaiki kelemahan algoritma C5.0. Hal ini memperjelas pentingnya strategi ensemble dalam klasifikasi teks yang kompleks. Temuan ini menunjukkan bahwa penggunaan SMOTE dan XGBoost secara signifikan dapat meningkatkan performa model, membantu dalam memahami persepsi publik secara lebih akurat.*

Kata Kunci *– Algoritma C5.0, SMOTE, XGBoost, Ensemble, Klasifikasi*

I. PENDAHULUAN

BPJS Kesehatan merupakan program jaminan kesehatan nasional yang bertujuan untuk memberikan akses layanan kesehatan yang merata bagi seluruh penduduk Indonesia[1]. Program ini telah mencatat peningkatan signifikan dalam jumlah peserta, mencapai hampir 250 juta jiwa hingga Januari 2023[2]. Namun, di balik peningkatan jumlah peserta, kualitas pelayanan BPJS Kesehatan sering menjadi sorotan karena berbagai keluhan dan ketidakpuasan yang diungkapkan oleh masyarakat[3]. Oleh karena itu, penting untuk menganalisis sentimen masyarakat terhadap layanan BPJS Kesehatan guna mendapatkan wawasan yang lebih mendalam mengenai persepsi mereka.

Sentimen masyarakat merupakan indikator penting yang dapat digunakan untuk menilai efektivitas layanan BPJS Kesehatan. Analisis sentimen menggunakan algoritma pembelajaran mesin (Machine Learning) memungkinkan identifikasi pola dan tren dalam data teks yang dapat memberikan gambaran yang lebih jelas mengenai kepuasan atau ketidakpuasan masyarakat[4]. Penelitian ini berfokus pada implementasi algoritma C5.0 untuk mengklasifikasikan sentimen masyarakat terhadap pelayanan BPJS Kesehatan di Indonesia. Algoritma C5.0 dipilih karena kemampuannya yang lebih baik dalam menangani data dengan variabel kategorikal yang kompleks dan memberikan interpretasi yang mudah dipahami[5][6].

Algoritma C5.0 adalah versi lanjutan dari algoritma C4.5 yang terkenal dalam pengklasifikasian pohon keputusan[7]. Keunggulan utama C5.0 dibandingkan C4.5 termasuk kecepatan pemrosesan yang lebih tinggi, efisiensi memori yang lebih baik, dan kemampuan menangani data yang hilang dengan lebih efektif [8]. Selain itu, C5.0 mendukung teknik Boosting yang dapat meningkatkan akurasi model dengan mengkombinasikan beberapa model sederhana menjadi satu model yang lebih baik[9]. Teknik ini sangat relevan untuk meningkatkan performa analisis sentimen pada dataset yang besar dan beragam seperti yang digunakan dalam penelitian ini. Penerapan algoritma C5.0 yang telah dioptimalkan dengan Boosting diharapkan hasil penelitian nantinya dapat memberikan wawasan yang lebih akurat dan mendalam tentang persepsi masyarakat terhadap layanan BPJS Kesehatan.

Data yang digunakan dalam penelitian ini berasal dari dataset "Kepuasan Masyarakat Terhadap Layanan BPJS Kesehatan di Indonesia tahun 2023-2024" yang tersedia di Kaggle, dengan jumlah data sebanyak 3060 entri. Dataset ini mencakup berbagai aspek layanan BPJS Kesehatan dan mencatat respon masyarakat dalam bentuk teks yang dapat dianalisis untuk menilai sentimen keseluruhan. Hasil analisis sentimen dapat digunakan oleh BPJS Kesehatan untuk memahami lebih baik kebutuhan dan harapan masyarakat, serta melakukan perbaikan layanan yang lebih terarah. Dengan demikian, penelitian ini tidak hanya memberikan pemahaman yang lebih baik tentang sentimen masyarakat tetapi juga membantu dalam pengambilan keputusan strategis untuk peningkatan layanan kesehatan di Indonesia.

II. PENELITIAN YANG TERKAIT

Penelitian sebelumnya menunjukkan bahwa meskipun algoritma C5.0 mampu menghasilkan akurasi tinggi dalam analisis hasil audit 5S, algoritma ini rentan terhadap overfitting dan tidak efektif dalam menangani data dengan dimensi tinggi atau tidak terstruktur seperti teks atau gambar[10]. Penelitian lainnya disimpulkan bahwa C5.0 memiliki kelemahan dalam menangani data tidak seimbang, sering kali mengabaikan kelas minoritas dan menekankan pentingnya preprocessing data yang kompleks[11]. Penelitian selanjutnya mengidentifikasi masalah independensi variabel dalam Naïve Bayes Classifier yang dapat mengurangi akurasi klasifikasi, meskipun C5.0 mengoptimalkan klasifikasi dengan mengidentifikasi variabel berpengaruh[12].

Hasil penelitian menyarankan peningkatan akurasi C5.0 melalui validasi silang, namun juga menghadapi keterbatasan dalam efisiensi memori dan kecepatan pada dataset besar[13]. Peneliti lain menemukan bahwa C5.0 efektif dalam mengatasi atribut multi-nilai dan yang hilang, tetapi memerlukan penentuan parameter yang tepat[14]. Peneliti lain menemukan bahwa C5.0 menghasilkan akurasi sangat baik dalam prediksi morfologi jamur dengan validasi split dan cross-validation, tetapi memerlukan penanganan khusus pada data bervariasi[15].

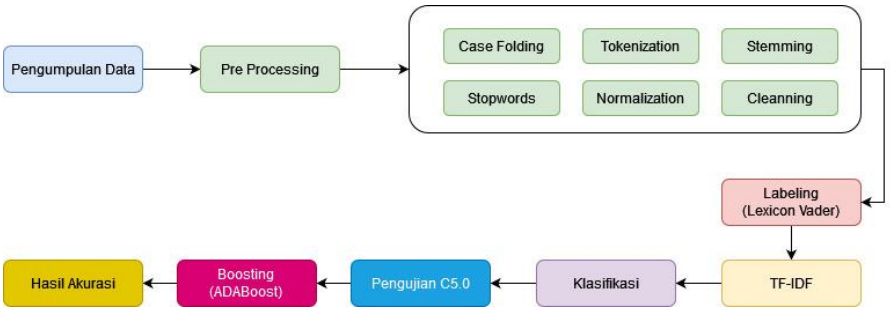
Peneliti menggunakan algoritma C5.0 untuk memprediksi persediaan buah pada UD. Bunda Syafira Buah, menghasilkan model yang baik tetapi membutuhkan variasi data yang lebih luas dan mempertimbangkan faktor eksternal seperti musim dan tren pasar[16]. Peneliti selanjutnya mengaplikasikan C5.0 untuk memprediksi kondisi kelahiran bayi, menghadapi kesulitan dalam perhitungan gain dan entropy serta kebutuhan validasi silang untuk meningkatkan akurasi[17].

Peneliti lain menerapkan C5.0 untuk prediksi kesesuaian lahan kedelai, menemukan kesulitan dalam menangani variabel lingkungan yang dinamis dan memerlukan data terstruktur[18]. Yulianti dan Syafrizal menggunakan C5.0 untuk mengevaluasi pemahaman siswa dalam pembelajaran online, menemukan bahwa hasil dapat dipengaruhi oleh variabel eksternal seperti kondisi jaringan dan perangkat yang digunakan[19], dan peneliti lain menggunakan algoritma C4.5 untuk mencatat kesulitan serupa dalam konteks pembelajaran daring yang dinamis dan heterogen[20].

Penelitian ini menawarkan keterbaruan dengan mengintegrasikan teknik boosting bersama algoritma C5.0 untuk mengatasi kelemahan yang ditemukan dalam penelitian sebelumnya, seperti masalah overfitting, kesulitan menangani data yang tidak seimbang, dan kebutuhan preprocessing data yang ekstensif. Dengan menerapkan teknik boosting, model dapat lebih baik menangani kelas minoritas dan meningkatkan generalisasi pada data baru. Pendekatan ini juga mencakup validasi silang untuk memastikan keandalan hasil, serta preprocessing data yang lebih komprehensif untuk memastikan data yang bersih dan siap digunakan dalam algoritma C5.0. Keterbaruan ini diharapkan dapat memberikan kontribusi signifikan dalam meningkatkan akurasi dan efektivitas model klasifikasi sentimen terhadap layanan BPJS Kesehatan, memberikan wawasan yang lebih mendalam tentang persepsi masyarakat, dan membantu dalam pengambilan keputusan yang lebih tepat untuk peningkatan layanan.

III. METODE PENELITIAN

Metode penelitian yang diilustrasikan dalam diagram terdiri dari beberapa langkah yang terstruktur untuk mengimplementasikan analisis sentimen terhadap layanan BPJS Kesehatan menggunakan algoritma C5.0 dan teknik Boosting. Berikut adalah gambar tahapan penelitian yang akan dilakukan:



Gbr 1. Metode Penelitian

Berikut adalah penjelasan rinci dari setiap tahapannya:

A. Pengumpulan Data

Langkah pertama adalah pengumpulan data dari dataset "Kepuasan Masyarakat Terhadap Layanan BPJS Kesehatan di Indonesia tahun 2023-2024" yang tersedia di Kaggle. Dataset ini berisi respon masyarakat terhadap layanan BPJS Kesehatan dalam bentuk teks yang perlu dianalisis untuk memahami sentimen mereka.

B. Pre-Processing

Data mentah yang dikumpulkan akan diproses lebih lanjut melalui beberapa tahap pre-processing untuk memastikan data siap digunakan dalam analisis. Tahap-tahap ini meliputi:

a. Case Folding: Mengubah semua huruf dalam teks menjadi huruf kecil untuk konsistensi. Ini penting untuk memastikan bahwa perbedaan antara huruf besar dan kecil tidak mempengaruhi analisis.
b. Tokenization: Memecah teks menjadi unit-unit kata atau token. Proses ini membantu dalam mengidentifikasi dan mengisolasi kata-kata individu dalam teks.
c. Stemming: Mengubah kata-kata ke bentuk dasarnya atau root word. Misalnya, kata "membeli", "membelikan", dan "pembeli" akan diubah menjadi "beli". Ini mengurangi kompleksitas data dan membantu dalam analisis yang lebih akurat.
d. Stopwords Removal: Menghilangkan kata-kata umum yang tidak memberikan makna signifikan, seperti "dan", "atau", "yang". Kata-kata ini tidak banyak berkontribusi pada sentimen dan bisa mengaburkan hasil analisis.
e. Normalization: Mengubah bentuk kata tidak baku atau slang ke bentuk baku. Misalnya, kata-kata seperti "gak" diubah menjadi "tidak". Ini penting untuk memastikan data konsisten.
f. Cleaning: Menghapus karakter atau simbol yang tidak diperlukan, seperti tanda baca yang tidak relevan atau URL dalam teks.

C. Labeling (Lexicon Vader)

Setelah data diproses, langkah berikutnya adalah pemberian label menggunakan metode Lexicon Vader. Metode ini digunakan untuk menentukan polaritas sentimen (positif, negatif, atau netral) berdasarkan teks

yang ada. Vader adalah salah satu pendekatan yang sering digunakan untuk analisis sentimen karena kemampuannya untuk menangani bahasa sehari-hari dan slang dengan baik.

D. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) digunakan untuk mengubah teks menjadi fitur numerik yang dapat digunakan oleh algoritma pembelajaran mesin. TF-IDF mengukur seberapa penting sebuah kata dalam dokumen relatif terhadap keseluruhan korpus. Ini membantu dalam mengidentifikasi kata-kata yang relevan untuk klasifikasi sentimen.

E. Klasifikasi

Data yang telah diubah menjadi fitur numerik kemudian diklasifikasikan menggunakan algoritma C5.0. Algoritma C5.0 adalah pengembangan dari algoritma C4.5 yang digunakan untuk membangun pohon keputusan. Keunggulan C5.0 meliputi kecepatan pemrosesan yang lebih tinggi, efisiensi memori yang lebih baik, dan kemampuan menangani data yang hilang.

F. Pengujian C5.0

Model yang dibangun dengan algoritma C5.0 diuji untuk mengevaluasi performanya dalam mengklasifikasikan sentimen. Pengujian ini melibatkan pengukuran metrik seperti akurasi, presisi, recall, dan F1-score untuk menilai sejauh mana model dapat mengklasifikasikan sentimen dengan benar. Algoritma C5.0 menawarkan sejumlah kelebihan signifikan dalam analisis sentimen dibandingkan dengan pendahulunya, C4.5. Salah satu keunggulan utama adalah kecepatan pemrosesan yang lebih tinggi dan efisiensi memori yang lebih baik, yang memungkinkan algoritma ini untuk menangani dataset besar dengan lebih cepat dan dengan penggunaan sumber daya yang lebih sedikit. C5.0 juga memiliki kemampuan yang lebih baik dalam menangani data yang hilang dan variabel kategorikal yang kompleks, membuatnya lebih fleksibel dalam berbagai situasi data nyata. Selain itu, algoritma ini mendukung teknik Boosting, seperti ADABoost, yang dapat meningkatkan akurasi model secara signifikan dengan menggabungkan beberapa model lemah menjadi satu model kuat. Keunggulan lainnya termasuk kemampuan interpretasi yang lebih mudah dan transparansi dalam pembuatan pohon keputusan, yang memudahkan pengguna untuk memahami dan menjelaskan hasil analisis. Kombinasi dari keunggulan-keunggulan ini membuat C5.0 menjadi pilihan yang sangat efektif untuk tugas-tugas klasifikasi, termasuk analisis sentimen terhadap layanan BPJS Kesehatan.

G. Boosting (ADABoost)

Teknik Boosting, khususnya ADABoost diterapkan untuk meningkatkan akurasi model. Boosting bekerja dengan mengkombinasikan beberapa model lemah menjadi satu model kuat. Dalam konteks ini, ADABoost digunakan untuk memperbaiki kesalahan yang dibuat oleh model C5.0, sehingga hasil akhirnya lebih akurat.

H. Hasil Akurasi

Akurasi dari model yang telah di-boosting dievaluasi dan dibandingkan dengan model awal. Evaluasi ini penting untuk memastikan bahwa teknik Boosting memberikan peningkatan yang signifikan dalam performa klasifikasi sentimen. Hasil akhir akan digunakan untuk membuat kesimpulan dan rekomendasi.

IV. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil dan diskusi yang bertujuan untuk menjawab pertanyaan penelitian terkait optimasi algoritma C5.0 dalam klasifikasi ulasan masyarakat terhadap layanan BPJS Kesehatan. Penelitian ini menggunakan teknik Boosting dan SMOTE untuk meningkatkan akurasi klasifikasi sentimen. Metode ini diharapkan dapat memberikan wawasan yang lebih baik tentang bagaimana publik merespons layanan BPJS Kesehatan berdasarkan data teks dari media sosial.

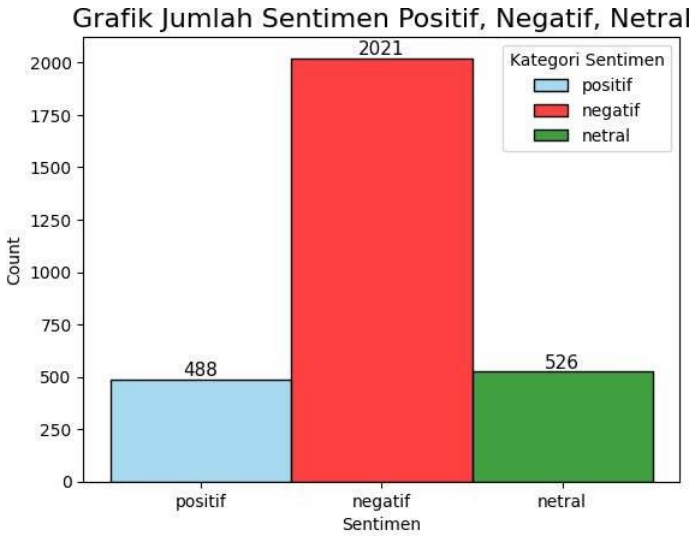
A. Tahap Preprocessing

Proses praproses data dilakukan untuk membersihkan dan mempersiapkan data ulasan masyarakat terhadap layanan BPJS Kesehatan agar siap digunakan dalam analisis. Tahapan ini melibatkan beberapa langkah penting,

dimulai dengan pengumpulan data dari sumber media sosial. Setelah data terkumpul, proses case folding dilakukan untuk mengubah seluruh teks menjadi huruf kecil, memastikan konsistensi dalam analisis teks. Kemudian, tokenisasi diterapkan untuk memecah teks menjadi unit-unit kata yang lebih mudah dianalisis. Selanjutnya, penghapusan stopwords dilakukan untuk mengeliminasi kata-kata umum yang tidak memiliki makna signifikan dalam analisis. Proses stemming mengubah kata-kata menjadi bentuk dasarnya, menyederhanakan teks, dan memudahkan algoritma dalam mengidentifikasi pola. Terakhir, proses normalisasi dan cleaning dilakukan untuk menghapus karakter khusus, simbol, angka, tautan, dan elemen lain yang tidak relevan, sehingga menghasilkan data yang lebih bersih dan terstruktur.

B. Labeling

Setelah melalui tahap praproses, data teks menjalani proses pelabelan menggunakan metode berbasis leksikon (Lexicon-Based) dengan menggunakan Lexicon Vader. Metode ini berfungsi untuk menentukan polaritas sentimen dari setiap teks, mengkategorikannya ke dalam sentimen positif, negatif, atau netral. Visualisasi distribusi sentimen menunjukkan adanya ketidakseimbangan kelas, di mana sentimen positif lebih dominan dibandingkan dengan sentimen negatif dan netral.



Gbr 2. Kategori Labeling

Grafik di atas menampilkan distribusi jumlah sentimen positif, negatif, dan netral dari ulasan masyarakat terhadap layanan BPJS Kesehatan. Dari hasil pelabelan, terlihat bahwa sentimen negatif mendominasi dengan total 2021 ulasan, menunjukkan bahwa mayoritas pengguna layanan memiliki persepsi yang kurang puas atau memiliki keluhan terhadap layanan tersebut. Sementara itu, sentimen netral terhitung sebanyak 526 ulasan, menunjukkan adanya sejumlah pengguna yang merasa layanan tersebut biasa saja atau tidak memiliki pendapat yang kuat. Sentimen positif yang hanya berjumlah 488 ulasan mengindikasikan bahwa hanya sebagian kecil dari pengguna yang merasa puas atau memberikan ulasan yang baik tentang layanan BPJS Kesehatan. Ketidakseimbangan dalam distribusi sentimen ini dapat memberikan informasi penting bagi penyedia layanan untuk memahami persepsi pengguna dan mengidentifikasi area yang perlu ditingkatkan guna meningkatkan kepuasan pelanggan.

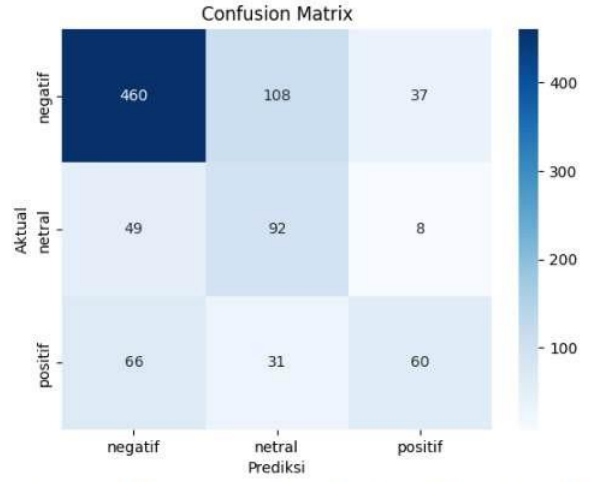
C. Pengujian dan Evaluasi Model
Algoritma C5.0

Pada bagian ini, dilakukan pengujian terhadap algoritma C5.0 untuk menilai kemampuannya dalam mengklasifikasikan sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan. Algoritma C5.0, yang merupakan pengembangan dari algoritma C4.5, dikenal sebagai salah satu metode yang efektif untuk membangun pohon keputusan dengan kecepatan pemrosesan yang lebih tinggi dan efisiensi memori yang lebih baik. Algoritma ini menghasilkan model klasifikasi yang dapat diinterpretasikan dengan mudah, sehingga memudahkan pengguna dalam memahami dan menjelaskan hasil analisis.

Algoritma C5.0 bekerja dengan membagi data menjadi subset berdasarkan atribut yang memaksimalkan gain ratio, yang merupakan ukuran seberapa baik atribut memisahkan kelas-kelas target. Gain ratio merupakan modifikasi dari information gain, dan digunakan untuk mengatasi kelemahan information gain dalam bias terhadap atribut dengan banyak nilai. Rumus untuk gain ratio adalah:

$$\text{Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}}$$

Di mana information gain dihitung sebagai perbedaan antara entropi sebelum dan sesudah pembagian berdasarkan atribut tertentu, dan split information merupakan ukuran entropi dari distribusi instans di antara berbagai nilai atribut. Dengan mengaplikasikan C5.0, diharapkan model yang dihasilkan mampu menangkap pola dalam data dan menghasilkan prediksi yang akurat tentang sentimen publik terhadap layanan BPJS Kesehatan.



Accuracy Score untuk Decision Tree Model (C5.0) :: 0.6717892425905598

	precision	recall	f1-score	support
negatif	0.80	0.76	0.78	605
netral	0.40	0.62	0.48	149
positif	0.57	0.38	0.46	157
accuracy			0.67	911
macro avg	0.59	0.59	0.57	911
weighted avg	0.69	0.67	0.68	911

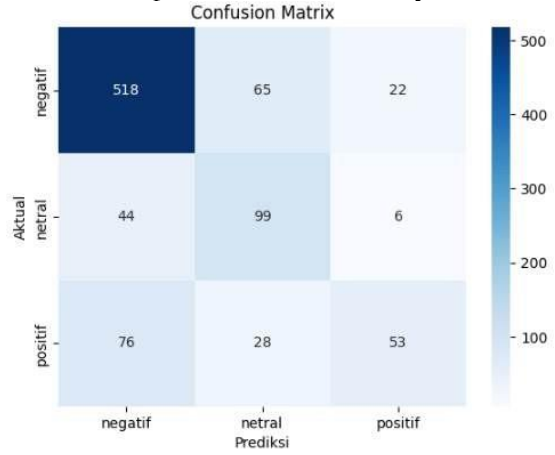
Gbr 3. Confusion Matrix Pengujian Algoritma C5.0

Berdasarkan confusion matrix dan metrik evaluasi yang dihasilkan dari pengujian model C5.0, terlihat bahwa model ini memiliki akurasi keseluruhan sebesar 67%. Sentimen negatif memiliki tingkat presisi dan recall yang relatif tinggi, yaitu masing-masing sebesar 80% dan 76%, menunjukkan bahwa model mampu mengidentifikasi ulasan negatif dengan baik. Namun, untuk sentimen netral dan positif, model menunjukkan performa yang lebih rendah, dengan presisi dan recall masing-masing sebesar 40% dan 62% untuk netral, serta 57% dan 38% untuk positif. F1-score yang lebih tinggi untuk sentimen negatif (78%) dibandingkan dengan netral (48%) dan positif (46%) mengindikasikan bahwa model lebih efektif dalam memprediksi ulasan negatif dibandingkan dengan yang lain. Ketidakseimbangan performa ini dapat disebabkan oleh distribusi data yang tidak seimbang, di mana jumlah ulasan negatif jauh lebih banyak dibandingkan dengan ulasan netral dan positif. Meskipun teknik SMOTE telah diterapkan, hasil ini menunjukkan bahwa perlu ada peningkatan lebih lanjut dalam teknik pemrosesan data atau pemilihan fitur untuk meningkatkan performa klasifikasi pada sentimen netral dan positif.

Algoritma C5.0 + ADABoost

Pada bagian ini, dilakukan pengujian kombinasi algoritma C5.0 dengan teknik Boosting menggunakan ADABoost untuk meningkatkan akurasi klasifikasi sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan. ADABoost adalah salah satu metode ensemble learning yang bertujuan untuk meningkatkan performa model dasar dengan menggabungkan beberapa model lemah menjadi satu model yang lebih kuat. Dalam konteks ini, ADABoost diterapkan untuk mengatasi kelemahan dari model C5.0 yang sebelumnya menunjukkan akurasi yang kurang memuaskan, terutama dalam memprediksi sentimen netral dan positif. Dengan menggunakan ADABoost, setiap iterasi pembelajaran memperbaiki kesalahan prediksi dari iterasi sebelumnya dengan memberikan bobot lebih pada

data yang salah klasifikasi, sehingga secara bertahap meningkatkan kemampuan model untuk memprediksi dengan lebih akurat. Pengujian ini diharapkan dapat menunjukkan peningkatan signifikan dalam metrik evaluasi seperti akurasi, presisi, recall, dan F1-score, yang pada akhirnya memberikan wawasan yang lebih baik mengenai sentimen publik terhadap layanan BPJS Kesehatan. Hasil pengujian dari kombinasi algoritma C5.0 dan ADABoost akan dianalisis berdasarkan matriks kebingungan yang ditampilkan di bawah ini, yang menunjukkan kinerja model dalam memprediksi kategori sentimen negatif, netral, dan positif dari ulasan masyarakat terhadap layanan BPJS Kesehatan.



Accuracy Score untuk AdaBoost Model :: 0.7354555433589463

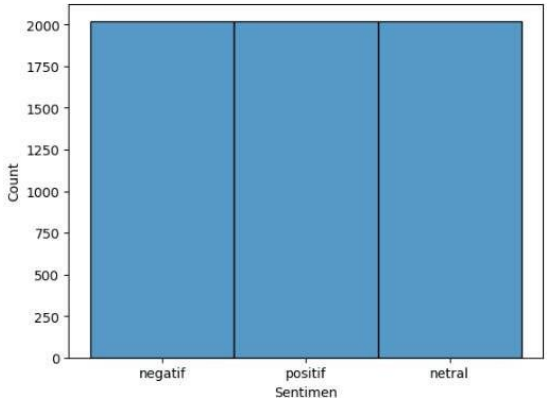
	precision	recall	f1-score	support
negatif	0.81	0.86	0.83	605
netral	0.52	0.66	0.58	149
positif	0.65	0.34	0.45	157
accuracy			0.74	911
macro avg	0.66	0.62	0.62	911
weighted avg	0.74	0.74	0.73	911

Gbr 4. Confusion Matrix Pengujian Algoritma C5.0 + ADABoost

Hasil pengujian menggunakan kombinasi algoritma C5.0 dan ADABoost menunjukkan peningkatan performa model dalam klasifikasi sentimen ulasan terhadap layanan BPJS Kesehatan, dengan akurasi keseluruhan sebesar 73.5%. Dalam matriks kebingungan, sentimen negatif memiliki presisi 81% dan recall 86%, yang berarti model sangat efektif dalam mengenali dan mengklasifikasikan ulasan negatif dengan benar. Untuk sentimen netral, model menunjukkan presisi 52% dan recall 66%, menunjukkan peningkatan kemampuan dalam mengklasifikasikan ulasan netral dibandingkan sebelumnya. Sentimen positif, meskipun masih memerlukan peningkatan, menunjukkan presisi 65% dengan recall 34%. Penggunaan ADABoost berhasil meningkatkan f1-score terutama pada kelas negatif dan netral, dengan nilai f1-score tertinggi pada kelas negatif sebesar 83%. Ini menandakan bahwa ADABoost membantu dalam memperbaiki prediksi dengan mengurangi kesalahan pada kelas mayoritas dan memberikan bobot lebih pada misclassifications yang berulang, sehingga secara keseluruhan meningkatkan kemampuan model untuk menangkap pola dalam data dengan lebih baik. Hasil ini menegaskan bahwa ADABoost efektif dalam meningkatkan keandalan model dalam menganalisis sentimen publik terhadap layanan BPJS Kesehatan.

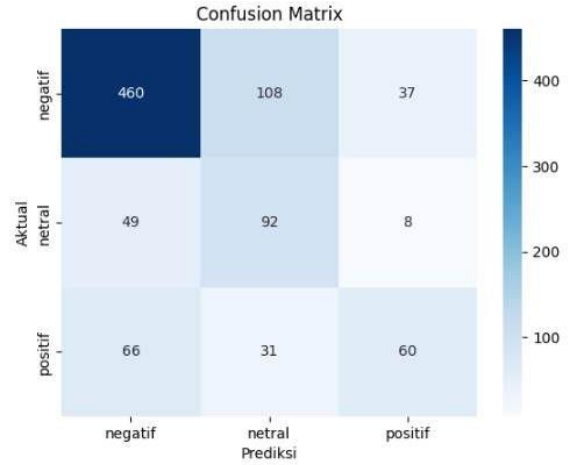
Algoritma C5.0 + SMOTE

Pada bagian ini, dilakukan pengujian algoritma C5.0 yang dioptimalkan dengan teknik SMOTE (Synthetic Minority Over-sampling Technique) untuk meningkatkan akurasi klasifikasi sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan. SMOTE digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset, di mana kelas minoritas mendapatkan representasi yang kurang dibandingkan dengan kelas mayoritas. Teknik ini bekerja dengan menciptakan contoh data sintetis untuk kelas minoritas dengan cara menginterpolasi antara contoh yang ada, sehingga memperbesar ukuran kelas minoritas dan menyeimbangkan distribusi data. Dengan demikian, SMOTE membantu model dalam mempelajari pola dari setiap kelas secara lebih efektif, mengurangi bias terhadap kelas mayoritas, dan meningkatkan performa model dalam mengklasifikasikan data. Hasil penerapan SMOTE pada dataset ini dapat dilihat pada grafik di bawah, yang menunjukkan distribusi kelas setelah penerapan teknik ini.



Gbr 5. Penyeimbangan Data Menggunakan SMOTE

Grafik di atas menunjukkan hasil penerapan teknik SMOTE pada dataset sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan. Setelah menggunakan SMOTE, distribusi data di antara kelas sentimen negatif, positif, dan netral menjadi seimbang, masing-masing memiliki jumlah sekitar 2000 sampel. Hal ini menunjukkan bahwa teknik SMOTE berhasil menyeimbangkan dataset yang sebelumnya didominasi oleh kelas negatif, dengan menambahkan sampel sintetis pada kelas positif dan netral yang sebelumnya lebih sedikit. Dengan distribusi yang seimbang ini, algoritma C5.0 diharapkan dapat mempelajari pola dari setiap kelas dengan lebih baik, mengurangi bias yang disebabkan oleh ketidakseimbangan kelas, dan meningkatkan akurasi prediksi untuk semua kategori sentimen. Sehingga, model yang dihasilkan akan lebih mampu memberikan representasi yang akurat dari persepsi publik terhadap layanan BPJS Kesehatan, serta meningkatkan validitas dan reliabilitas hasil analisis sentimen secara keseluruhan.



Accuracy Score untuk Decision Tree Model (C5.0) :: 0.6717892425905598

	precision	recall	f1-score	support
negatif	0.80	0.76	0.78	605
netral	0.40	0.62	0.48	149
positif	0.57	0.38	0.46	157
accuracy			0.67	911
macro avg	0.59	0.59	0.57	911
weighted avg	0.69	0.67	0.68	911

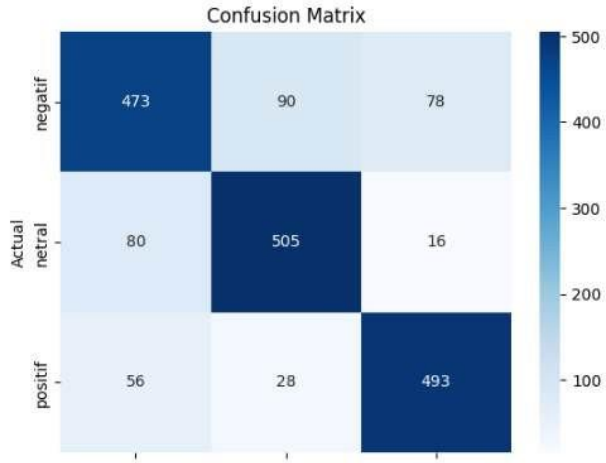
Gbr 6. Confusion Matrix Pengujian Algoritma C5.0 + SMOTE

Hasil pengujian algoritma C5.0 yang dioptimalkan dengan SMOTE menunjukkan akurasi keseluruhan sebesar 67.2%. Meskipun SMOTE berhasil menyeimbangkan distribusi data antar kelas sentimen, model masih menunjukkan performa yang lebih baik dalam mengklasifikasikan ulasan negatif dibandingkan dengan netral dan positif. Presisi dan recall untuk sentimen negatif masing-masing adalah 80% dan 76%, yang menunjukkan bahwa model cukup efektif dalam mengenali ulasan negatif. Namun, untuk sentimen netral dan positif, presisi dan recall relatif lebih rendah, dengan masing-masing mencapai 40% dan 62% untuk netral, serta 57% dan 38% untuk positif. Ketidakseimbangan ini mungkin disebabkan oleh kompleksitas konteks dalam ulasan netral dan positif, yang mungkin memiliki fitur yang lebih ambigu atau kurang berbeda dibandingkan ulasan negatif, sehingga menyulitkan model untuk memprediksi dengan akurat. Meskipun SMOTE meningkatkan representasi kelas minoritas, tantangan yang lebih besar dalam memahami konteks teks dan fitur yang kurang informatif mungkin masih mempengaruhi

kemampuan model untuk memprediksi dengan lebih tepat. Selain itu, kemungkinan juga terdapat noise dalam data atau fitur yang kurang relevan yang dapat mempengaruhi akurasi model. Untuk meningkatkan performa, pendekatan lain seperti pemilihan fitur yang lebih baik atau penggunaan teknik tambahan untuk menangani ketidakseimbangan mungkin diperlukan.

Algoritma C5.0 + SMOTE + ADABoost

Pada bagian ini, dilakukan pengujian dengan kombinasi algoritma C5.0, SMOTE, dan ADABoost untuk meningkatkan performa model dalam klasifikasi sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan. Kombinasi ini bertujuan untuk mengoptimalkan keakuratan prediksi dengan mengatasi ketidakseimbangan kelas melalui SMOTE dan memperkuat kemampuan klasifikasi C5.0 menggunakan teknik Boosting dengan ADABoost. SMOTE membantu menyeimbangkan distribusi kelas dengan menambah sampel sintetis untuk kelas minoritas, sementara ADABoost meningkatkan akurasi model dengan menggabungkan beberapa model lemah menjadi satu model yang lebih kuat. Melalui pengujian ini, diharapkan peningkatan yang signifikan dalam akurasi, presisi, recall, dan f1-score dibandingkan dengan penggunaan algoritma secara terpisah. Hasil pengujian dan analisis kinerja model dapat dilihat pada grafik di bawah ini.



	precision	recall	f1-score	support
0	0.78	0.74	0.76	641
1	0.81	0.84	0.83	601
2	0.84	0.85	0.85	577
accuracy			0.81	1819
macro avg	0.81	0.81	0.81	1819
weighted avg	0.81	0.81	0.81	1819

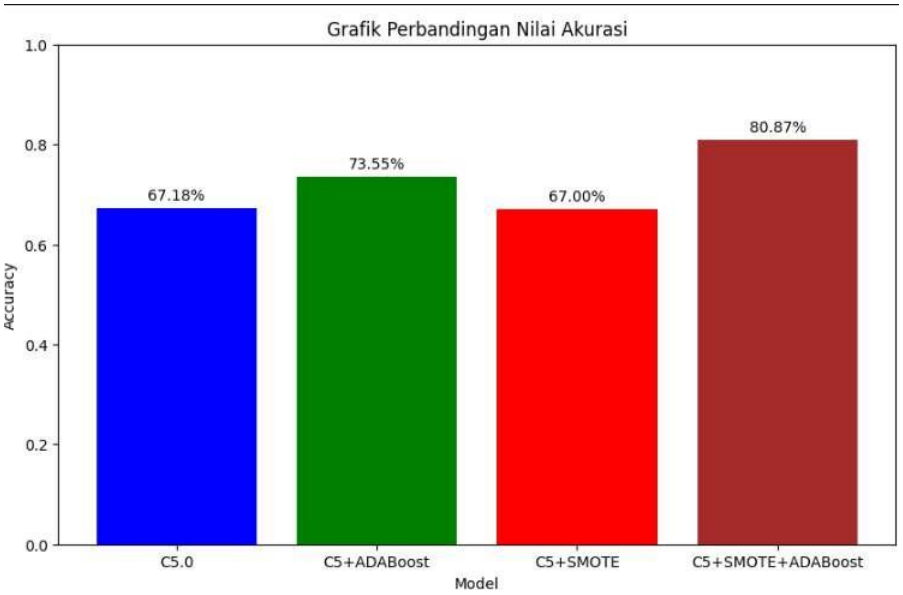
C5.0+SMOTE+XGBoost Accuracy: 0.8086860912589334

Gbr 7. Confusion Matrix Pengujian Algoritma C5.0 + SMOTE + ADABoost

Hasil pengujian kombinasi algoritma C5.0, SMOTE, dan ADABoost menunjukkan peningkatan signifikan dalam performa model dengan akurasi keseluruhan mencapai 80.9%. Matriks kebingungan menunjukkan bahwa model ini mampu mengklasifikasikan sentimen negatif, netral, dan positif dengan lebih baik dibandingkan pengujian sebelumnya. Sentimen negatif memiliki presisi 78% dan recall 74%, sedangkan sentimen netral menunjukkan presisi 81% dan recall 84%. Untuk sentimen positif, model mencapai presisi 84% dan recall 85%, yang merupakan peningkatan signifikan dari hasil sebelumnya. Teknik ADABoost terbukti efektif dalam meningkatkan akurasi dengan memperkuat model dasar melalui iterasi dan penyesuaian bobot pada kesalahan prediksi, sehingga memperbaiki kemampuan model dalam mengklasifikasikan data yang lebih sulit. Penggunaan SMOTE memastikan distribusi data yang seimbang, sementara ADABoost meningkatkan keakuratan dengan mengurangi bias dan meningkatkan ketahanan model terhadap variabilitas data. Kombinasi ini menunjukkan bahwa ADABoost mampu mengatasi kelemahan dari model C5.0 yang sebelumnya memiliki akurasi lebih rendah, memberikan hasil yang lebih andal dan akurat dalam analisis sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan.

D. Hasil Komparasi

Pada bagian ini, dilakukan analisis komparatif untuk mengevaluasi kinerja berbagai kombinasi metode pengujian yang telah diterapkan sebelumnya, termasuk C5.0, SMOTE, dan ADABoost, dalam mengklasifikasikan sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan. Analisis ini bertujuan untuk mengidentifikasi pendekatan terbaik yang mampu memberikan akurasi tertinggi dan meningkatkan kemampuan model dalam memahami pola data. Setiap kombinasi metode memiliki keunggulan dan kelemahan tersendiri, dan hasil komparasi ini akan memberikan wawasan tentang efektivitas masing-masing pendekatan dalam menangani ketidakseimbangan data dan kompleksitas fitur teks. Hasil komparasi kinerja model dapat dilihat pada grafik di bawah ini, yang menunjukkan perbedaan akurasi, presisi, recall, dan f1-score dari setiap metode pengujian yang digunakan.



Gbr 8. Grafik Komparasi Model

Grafik perbandingan nilai akurasi di atas menggambarkan hasil komparasi dari beberapa kombinasi metode pengujian yang telah diterapkan, yaitu C5.0, C5.0 dengan ADABoost, C5.0 dengan SMOTE, dan kombinasi C5.0, SMOTE, dan ADABoost. Algoritma C5.0 dasar memiliki akurasi sebesar 67.18%, menunjukkan bahwa algoritma ini memiliki kemampuan dasar yang cukup baik dalam klasifikasi sentimen. Penerapan ADABoost pada C5.0 meningkatkan akurasi menjadi 73.55%, yang menandakan bahwa teknik Boosting efektif dalam memperbaiki kesalahan prediksi dari model dasar. Kombinasi C5.0 dan SMOTE, meskipun bertujuan untuk menyeimbangkan distribusi kelas, tidak meningkatkan akurasi secara signifikan, dengan hasil akurasi 67.00%. Namun, kombinasi optimal dari C5.0, SMOTE, dan ADABoost mencapai akurasi tertinggi sebesar 80.87%, menunjukkan bahwa penggunaan bersama teknik SMOTE dan ADABoost berhasil memaksimalkan performa model dengan menangani ketidakseimbangan kelas dan meningkatkan kekuatan prediktif model. Ini mengindikasikan bahwa kombinasi teknik yang tepat dapat secara signifikan meningkatkan akurasi dalam klasifikasi sentimen ulasan masyarakat terhadap layanan BPJS Kesehatan.

V. KESIMPULAN

Penelitian ini telah mengeksplorasi berbagai kombinasi algoritma dan teknik untuk meningkatkan akurasi dalam klasifikasi sentimen terkait pemindahan Ibu Kota Nusantara (IKN) di Indonesia. Kombinasi algoritma yang digunakan meliputi Naive Bayes, Random Forest, SMOTE, dan XGBoost. Hasil penelitian menunjukkan bahwa kombinasi Random Forest dengan SMOTE memberikan hasil terbaik dengan akurasi mencapai 91,25%. Penerapan SMOTE berhasil menyeimbangkan dataset yang tidak seimbang, meningkatkan kemampuan model dalam mendeteksi sentimen dari kelas minoritas. Dalam hasil komparasi, model Naive Bayes (NB) memiliki akurasi 77,45%, yang meningkat menjadi 84,82% dengan XGBoost, dan 78,88% dengan SMOTE. Kombinasi NB dengan SMOTE dan XGBoost mencapai 85,97%. Model Random Forest (RF) sendiri mencapai 87,46% dan meningkat menjadi 88,78% dengan XGBoost. Menariknya, Random Forest dengan SMOTE menunjukkan akurasi tertinggi yaitu 91,25% dibandingkan dengan kombinasi RF, SMOTE, dan XGBoost yang memiliki akurasi 90,92%, karena penambahan XGBoost dapat menambah kompleksitas yang sedikit mengurangi kemampuan generalisasi model.

Peluang penelitian selanjutnya dapat difokuskan pada pengembangan model yang lebih adaptif dengan memasukkan teknik lain seperti deep learning atau penggunaan model transformer untuk menangani data teks yang lebih kompleks. Peneliti dapat melakukan tuning parameter lebih lanjut untuk Random Forest dan XGBoost guna mengoptimalkan kinerja model. Selain itu, studi masa depan dapat memperluas cakupan data dengan menggabungkan data dari berbagai platform media sosial atau menggunakan data real-time untuk menangkap dinamika sentimen yang terus berubah. Dengan pendekatan yang lebih holistik, penelitian mendatang dapat memberikan wawasan lebih dalam mengenai persepsi publik dan membantu dalam perumusan kebijakan yang lebih efektif, serta mengidentifikasi faktor-faktor lain yang dapat mempengaruhi efektivitas model prediktif.

UCAPAN TERIMA KASIH

Saya ingin mengucapkan terima kasih yang sebesar-besarnya kepada Universitas Hang Tuah Pekanbaru atas dukungan dan kontribusinya melalui pendanaan penelitian dan pengabdian kepada masyarakat tahun 2024.

DAFTAR PUSTAKA

[1] T. H. W. L. Deysi Liem Fat Salim, Nontje Rimbing, "Aksesibilitas Pembiayaan Kesehatan Dalam Program Jaminan Kesehatan Nasional," *Lex Soc.*, vol. VIII, no. 4, pp. 104–114, 2020.

[2] C. M. Annur, "Jumlah Peserta JKN BPJS Kesehatan Hampir Tembus 250 Juta Orang per Januari 2023," *Katadaa Media Network*, 2023. https://databoks.katadata.co.id/datapublish/2023/02/24/jumlah-peserta-jkn-bpjs-kesehatan-hampir-tembus-250-juta-orang-per-januari-2023

[3] R. Dewi and F. Jihad, "Hubungan Kualitas Pelayanan Kesehatan Dengan Kepuasan Pasien Rawat Jalan Pesertsa BPJS Kesehatan," *J. Kesehat. Tambusai*, vol. 4, no. 3, pp. 3662–3671, 2023, [Online]. Available: https://journal.universitaspahlawan.ac.id/index.php/jkt/article/view/16977

[4] Syahril Dwi Prasetyo, Shofa Shofiah Hilabi, and Fitri Nurapriani, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," *J. KomtekInfo*, vol. 10, pp. 1–7, 2023, doi: 10.35134/komtekinfo.v10i1.330.

[5] I. Berutu, "Penerapan Metode C5.0 Untuk Pengelompokkan Potensi Nasabah PT.Pegadaian Berdasarkan Pola Pembayaran Angsuran," *RESOLUSI Rekayasa Tek. Inform. dan Inf.*, vol. 1, no. 4, pp. 232–240, 2021, [Online]. Available: https://djournals.com/resolusi

[6] Z. Guo, Y. Shi, F. Huang, X. Fan, and J. Huang, "Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management," *Geosci. Front.*, vol. 12, no. 6, p. 101249, 2021, doi: 10.1016/j.gsf.2021.101249.

[7] Tanti, "Random Oversampling , Chi-Square , dan AdaBoost dalam Penanganan Ketidakseimbangan Kelas pada Klasifikasi C5 . 0," *J. Media Inform. Budidarma*, vol. 7, no. April, pp. 714–725, 2023, doi: 10.30865/mib.v7i2.5862.

[8] N. Benediktus and R. S. Oetama, "Algoritma Klasifikasi Decision Tree C5.0 untuk Memprediksi Performa Akademik Siswa Natanael," *Ultim. J. Tek. Inform.*, vol. 12, no. 1, pp. 14–19, 2020, [Online]. Available: https://www.kaggle.com/aljarah/xAPI-Edu-Data

[9] A. F. Siska and R. E. Putra, "Klasifikasi Tingkat Kepuasan Wali Murid Terhadap Hasil Belajar Anak Menggunakan Algoritma C5.0," *J. Informatics Comput. Sci.*, vol. 04, no. 04, pp. 432–435, 2023.

[10] C. L. Fantasy, F. L. M. Simanjuntak, R. L. A. Purba, Andrean, and O. Sihombing, "Analisis Komparasi Algoritma C5.0 Dan Naive Bayes Penentuan Penerima Beasiswa Universitas Prima Indonesia," *J. TEKINKOM*, vol. 6, no. 2, pp. 508–517, 2023, doi: 10.37600/tekinkom.v6i2.926.

[11] N. M. Asih, J. H. Jaman, and Y. Umaidah, "Analisis Sentimen Terhadap Bantuan Kuota Internet Dari Kemendikbud Dimasa Covid-19 Menggunakan Algoritma C5.0," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 5, no. 2, pp. 1–9, 2022, doi: 10.31539/intecoms.v5i2.2793.

[12] A. A. Wulandari, D. Retno, and S. Saputro, "KLASIFIKASI DATA MINING MENGGUNAKAN NAÏVE BAYES CLASSIFIER DENGAN ALGORITMA C5.0 (Classification Data Mining using Naïve Bayes Classifier with C5.0 Algorithm)," *Semin. Nas. Mat. Geom. Stat. dan Komputasi SeNa-MaGeStiK 2022*, vol. 0, no. 1, pp. 1–6, 2022, [Online]. Available: https://magestic.unej.ac.id/

[13] I. Aliyudin and A. P. Wahyu, "APPLICATION OF THE C5.0 ALGORITHM TO DETERMINE GOOD OR BAD ON 5S AUDIT RESULTS," *J. DARMA AGUNG*, vol. 30, no. 3, pp. 406–413, 2022.

[14] A. F. Cahyaningrum, Y. H. Chrisnanto, and A. K. Ningsih, "Enrichment : Journal of Multidisciplinary Research and Development Classification of Sentiment Towards BPJS Services Using the C50 Algorithm," vol. 1, no. 8, pp. 484–491, 2023.

[15] L. Karlitasari, I. W. Sriyasa, I. Wahyudi, and H. B. Santosi, "Prediksi Morfologi Jamur Menggunakan Algoritma C5.0," *J. Teknoinfo*, vol. 17, no. 1, p. 271, 2023, doi: 10.33365/jti.v17i1.2372.

[16] R. Ella Sari, Solikhun, and F. Rizky, "Penerapan Algoritma C5.0 dalam Memprediksi Persediaan Buah pada UD. Bunda Syafira Buah," *JUKI J. Komput. dan Inform.*, vol. 3, no. 2, pp. 59–63, 2021, doi: 10.53842/juki.v3i2.62.

[17] Siti Sahara Nasution, Natalia Silalahi, and Abdul Karim, "Implementasi Algoritma C5.0 Untuk Memprediksi Kondisi Kelahiran Bayi," *Bull. Comput. Sci. Res.*, vol. 2, no. 1, pp. 18–24, 2021, doi: 10.47065/bulletincsr.v2i1.138.

[18] N. Pratama and L. Andraini, "Model Prediksi Kesesuaian Lahan Kedelai Menggunakan C5. 0 Algoritma," *J. Portal Data*, vol. 2, no. 10, pp. 1–13, 2022, [Online]. Available: http://portaldata.org/index.php/portaldata/article/view/250

[19] A. Priyatna and Sanwani, "Evaluasi Pemahaman Siswa Dalam Proses Belajar Secara Online dengan Menggunakan Algoritma C5.0," *Resolusi Rekayasa Tek. Inform. dan Inf.*, vol. 4, no. 3, pp. 300–309, 2024.

[20] A. Ramadhan, "Sistem Pendukung Keputusan Evaluasi Problematika Pendampingan Pembelajaran Daring dengan Algoritma C4.5," *J. Sistim Inf. dan Teknol.*, vol. 4, pp. 58–63, 2022, doi: 10.37034/jsisfotek.v4i2.124.