Jurnal Pendidikan dan Teknologi Indonesia (JPTI)
Vol. 5, No. 8, Agustus 2025, Hal. 2295-2318
DOI: https://doi.org/10.52436/1.jpti.880
p-ISSN: 2775-4227
e-ISSN: 2775-4219

Evaluasi Model Random Forest dengan Teknik SMOTE dan RUS untuk Klasifikasi Kerusakan Motor di Bengkel PLAVIX

Zacky Fahd Annahdli*[1], Elkaf Rahmawan Pramudya[2]

[1,2]Ilmu Komputer, Teknik Informatika, Universitas Dian Nuswantoro, Semarang, Indonesia
Email: [1]111202113422@mhs.dinus.ac.id, [2]elkaf.rahmawan@dsn.dinus.ac.id

Abstrak

Diagnosis kerusakan motor di bengkel konvensional masih sangat bergantung pada intuisi dan pengalaman subjektif mekanik, yang dapat menyebabkan inkonsistensi dan potensi kesalahan dalam penanganan kendaraan. Untuk mengatasi permasalahan tersebut, penelitian ini mengevaluasi kinerja algoritma *Random Forest* dalam mengklasifikasikan jenis kerusakan motor menggunakan dataset dari Bengkel PLAVIX. Data gejala diolah menggunakan *Term Frequency-Inverse Document Frequency* (TF-IDF), sedangkan fitur kategorikal dienkode dengan *Label Encoding*. Ketidakseimbangan data ditangani menggunakan *Synthetic Minority Over-sampling Technique* (SMOTE) dan *Random Under-Sampling* (RUS). Hasil menunjukkan bahwa *Random Forest* dengan SMOTE mampu meningkatkan akurasi dari 72,73% menjadi 77,27%, dengan peningkatan signifikan pada *presisi* sebesar 67,42%, serta *recall* sebesar 77,27% dan *F1-Score* sebesar 70,91% . Kombinasi SMOTE dan RUS juga memberikan keseimbangan yang lebih baik antara presisi dan recall. Studi ini membuktikan bahwa pendekatan *machine learning* dapat meningkatkan akurasi dan objektivitas diagnosis kerusakan motor, serta membantu bengkel dalam memberikan layanan perawatan kendaraan yang lebih andal dan efisien.

Kata Kunci: *machine learning, random forest, klasifikasi kerusakan motor, TF-IDF, SMOTE.*

Performance of Random Forest Model in Motorcycle Malfunctions Classification Based on PLAVIX Garage Data

Abstract

Diagnosis of motorcycle damage in conventional workshops still relies heavily on the intuition and subjective experience of mechanics, which can lead to inconsistencies and potential errors in vehicle handling. To overcome these problems, this study evaluates the performance of the Random Forest algorithm in classifying motorcycle damage types using a dataset from PLAVIX Workshop. Symptom data is processed using Term Frequency-Inverse Document Frequency (TF-IDF), while categorical features are encoded with Label Encoding. Data imbalance is handled using Synthetic Minority Over-sampling Technique (SMOTE) and Random Under-Sampling (RUS). Results showed that Random Forest with SMOTE was able to improve accuracy from 72.73% to 77.27%, with significant improvements in precision by 67.42%, as well as recall by 77.27% and F1-Score by 70.91%. The combination of SMOTE and RUS also provides a better balance between precision and recall. This study proves that machine learning approaches can improve the accuracy and objectivity of motorcycle fault diagnosis, and help repair shops provide more reliable and efficient vehicle maintenance services.

Keywords*: machine learning, random forest, classification of motor malfunctions, TF-IDF, SMOTE*

1. PENDAHULUAN

Dengan pertumbuhan populasi global yang terus meningkat, industri otomotif terus berkembang [1]. Sepeda motor, sebagai salah satu moda transportasi utama, memiliki peran penting dalam mobilitas masyarakat, terutama di negara berkembang. Berdasarkan data dari Badan Pusat Statistik (BPS), jumlah sepeda motor di Indonesia mencapai 125.305.332 unit pada tahun 2022, menunjukkan peningkatan signifikan dibanding tahun-tahun sebelumnya. Pertumbuhan ini membawa tantangan baru dalam hal perawatan dan perbaikan kendaraan, khususnya dalam mendeteksi dan mendiagnosis kerusakan secara efisien dan akurat.

Kerusakan motor merupakan permasalahan yang sering terjadi dalam dunia otomotif dan industri [2]–[4]. Identifikasi dini terhadap jenis kerusakan dapat membantu dalam perawatan dan perbaikan yang lebih efisien[5]. Namun, tantangan utama dalam klasifikasi kerusakan motor adalah ketidakseimbangan data, di mana beberapa

p-ISSN: 2775-4227
e-ISSN: 2775-4219

jenis kerusakan lebih jarang terjadi dibandingkan yang lain. Metode diagnosis manual yang umum digunakan di bengkel sangat bergantung pada pengalaman dan intuisi mekanik. Proses ini memiliki beberapa kelemahan, seperti subjektivitas dalam penilaian, waktu yang cukup lama, serta potensi kesalahan diagnosis yang tinggi. Kesalahan dalam diagnosis dapat menyebabkan perbaikan yang tidak tepat, meningkatkan biaya perbaikan, serta menurunkan kepercayaan pelanggan terhadap layanan bengkel [6]. Oleh karena itu, diperlukan solusi yang lebih akurat dan efisien untuk mengatasi permasalahan ini.

Layanan perawatan kendaraan yang baik bergantung pada diagnosis kerusakan motor yang akurat. Dengan diagnosis yang tepat, jenis kerusakan dapat diidentifikasi sejak dini, yang memungkinkan perbaikan yang efektif, tepat sasaran, dan murah. Selain itu, diagnosis yang tepat sangat penting untuk keselamatan pengendara karena kesalahan dalam mengidentifikasi kerusakan dapat menyebabkan kerusakan komponen vital saat kendaraan digunakan. Dari sudut pandang bisnis, diagnosis yang akurat meningkatkan kepercayaan pelanggan terhadap bengkel, meningkatkan reputasi layanan, dan mengurangi jumlah keluhan yang timbul karena perbaikan yang tidak efektif. Akibatnya, ada kemungkinan bahwa sistem diagnosis otomatis yang berbasis teknologi seperti pembelajaran mesin dapat membantu meningkatkan akurasi, konsistensi, dan efisiensi proses identifikasi kerusakan motor.

Teknologi kecerdasan buatan (AI) telah berkembang pesat dalam beberapa tahun terakhir, menghasilkan berbagai solusi inovatif, termasuk dalam industri otomotif [7]. Salah satu teknik AI yang banyak digunakan adalah machine learning (ML), yang memungkinkan sistem untuk belajar dari data historis dan membuat prediksi yang lebih akurat [7]. Dalam konteks perawatan kendaraan, ML dapat digunakan untuk menganalisis data dari berbagai sumber, seperti riwayat perbaikan bengkel, sensor kendaraan, serta data pengguna, guna mengidentifikasi pola-pola kerusakan yang umum terjadi.

Penelitian ini menggunakan dataset dari Bengkel PLAVIX yang berisi deskripsi kerusakan sepeda motor dalam bentuk teks dengan jumlah 110 data. Dataset ini memiliki beberapa tantangan utama, yaitu ketidakseimbangan kelas (imbalanced data), keberagaman fitur, serta kompleksitas pola hubungan antar fitur. Ketidakseimbangan data menjadi tantangan signifikan karena beberapa jenis kerusakan lebih sering terjadi dibandingkan yang lain, menyebabkan model ML cenderung bias terhadap kelas mayoritas [8]. Selain itu, keberagaman fitur dalam dataset ini mencakup kombinasi data kategorikal (merek, model, jenis motor) dan data berbasis teks (deskripsi gejala kerusakan) serta jenis kerusakan sebagai variabel target, sehingga membutuhkan teknik khusus untuk pemrosesan data seperti TF-IDF untuk menangani data teks pada fitur gejala dan encoding fitur kategorikal menggunakan LabelEncoder dari library scikit-learn [9], [10]. Hal ini menyebabkan model pembelajaran mesin cenderung bias terhadap kelas mayoritas.

Algoritma Random Forest (RF) diterapkan dalam penelitian dengan pendekatan yang beragam, tergantung pada komponen tambahan yang digunakan. Pendekatan pertama adalah RF standar yang diterapkan tanpa modifikasi tambahan, di mana prosesnya mencakup penggunaan dataset tertentu, penerapan prosedur standar seperti pemilihan fitur dan pembentukan pohon keputusan, serta analisis hasil berdasarkan metrik evaluasi seperti akurasi, presisi, dan recall. Pendekatan kedua mengintegrasikan RF dengan analisis data tambahan, seperti eksplorasi distribusi fitur dan korelasi antar variabel, yang bertujuan untuk memahami karakteristik dataset sebelum model dilatih. Sementara itu, pendekatan ketiga mengombinasikan RF dengan teknik prepocessing teks dan penyeimbangan data. Perpocessing teks mencakup normalisasi, stemming, penghapusan stopword, serta representasi teks menggunakan TF-IDF atau word embeddings, sedangkan teknik penyeimbangan data seperti SMOTE atau undersampling diterapkan untuk mengatasi ketimpangan distribusi kelas dalam dataset. Hal ini menyebabkan model pembelajaran mesin cenderung bias terhadap kelas mayoritas.

Beberapa penelitian sebelumnya telah menguji efektivitas metode ini pada berbagai jenis dataset, sebagaimana dirangkum dalam Tabel 1. Studi oleh Saputro & Rosiyadi (2022) menunjukkan bahwa metode RF + ROUS meningkatkan akurasi dari 97.5% menjadi 98%[11]. Dalam klasifikasi berbasis teks, penelitian Istiqamah & Rijal (2024) serta Aryanti dkk. (2023) menunjukkan bahwa RF + SMOTE meningkatkan akurasi dari 86% menjadi 88%, sedangkan RF + SMOTE + RUS dalam studi Istiqamah & Rijal (2024) meningkatkan akurasi dari 73% menjadi 75%[12][13]. Rincian lebih lanjut dapat dilihat dalam Tabel 1.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tabel 1 Perbandingan Kinerja Random Forest

Referensi	Model	Dataset	TF-IDF	Akurasi	Recall	Presisi	Kelebihan	Kekurangan
Istiqamah & Rijal (2024)	RF	Ulasan Konsumen (Teks)	Ya	73%	73%	98%	RF memberikan hasil yang cukup baik dalam klasifikasi ulasan	Kurang efektif dalam menangani ketidakseimbangan data
	RF+SMOTE			72%	75%	89%	SMOTE meningkatkan pengenalan kelas minoritas	Tidak memperhitungkan kelas mayoritas
	RF+SMOTE +RUS			75%	78%	89%	Kombinasi oversampling dan undersampling meningkatkan akurasi lebih baik	Risiko overfitting
Aryanti dkk. (2023)	RF	Ulasan Aplikasi Primaku (Teks)	Ya	86%	82%	84%	RF bekerja dengan baik dalam analisis sentimen aplikasi	Masih ada ketidakseimbangan data yang mempengaruhi hasil
	RF+SMOTE			88%	86%	85%	Meningkatkan keseimbangan kelas dan akurasi model	Tidak dibandingkan dengan metode SMOTE+RUS
Saputro & Rosiyadi (2022)	RF	Dataset Diabetes (Tabular)	Tidak	97.5%	-	-	Akurasi dasar sudah tinggi tanpa resampling	Dataset lebih kecil dibanding referensi lain
	RF+ROUS (SMOTE +RUS)			98%	-	-	ROUS sedikit meningkatkan akurasi	Peningkatan kecil karena dataset sudah relatif seimbang

Dari pendekatan tersebut, metode RF yang lebih komprehensif dipilih karena beberapa faktor utama. Preprocessing data meningkatkan kualitas fitur dengan menghilangkan noise dan informasi yang tidak relevan. Teknik penyeimbangan data mencegah bias terhadap kelas mayoritas, sehingga meningkatkan akurasi prediksi pada semua kelas. Analisis data tambahan memberikan pemahaman lebih mendalam tentang karakteristik dataset, yang mendukung pemilihan parameter optimal. Kombinasi preprocessing, penyeimbangan data, dan analisis data memungkinkan model mencapai akurasi lebih tinggi serta mengurangi risiko overfitting, sehingga meningkatkan generalisasi terhadap data baru.

Dalam penelitian ini, algoritma Random Forest (RF) dipilih karena beberapa alasan. Pertama, Random Forest dapat menangani klasifikasi data yang tidak seimbang [14] . Kedua, RF kompatibel dengan fitur kategorikal dan teks karena dapat bekerja dengan hasil ekstraksi teks menggunakan TF-IDF [15]. Ketiga, RF dikenal tahan terhadap overfitting karena menggabungkan prediksi dari berbagai pohon keputusan yang dilatih secara independen, sehingga menghasilkan model yang lebih generalis. Dengan menggabungkan keluaran dari banyak pohon keputusan, RF mengurangi risiko overfitting dan menawarkan akurasi yang lebih tinggi dalam prediksi [16]. Untuk menangani fitur berbasis teks dalam dataset ini, digunakan beberapa teknik Natural Language Processing (NLP), seperti normalisasi teks, stopword removal, stemming dengan Sastrawi, dan TF-IDF vectorization. Teknik ini bertujuan untuk meningkatkan kualitas data sebelum masuk ke dalam model ML [9].

Dalam eksplorasi data (EDA), terdapat tiga tahapan utama sebelum menerapkan model pembelajaran mesin, yaitu pembersihan data, penyeimbangan data, dan pemilihan fitur. Pembersihan data dilakukan untuk menghilangkan data yang tidak valid, duplikasi, serta menangani nilai yang hilang melalui imputasi, penghapusan outlier, dan transformasi data. Penyeimbangan data bertujuan untuk mengatasi ketidakseimbangan kelas dalam dataset klasifikasi menggunakan metode seperti SMOTE, yang menambahkan sampel sintetis untuk kelas minoritas, serta RUS, yang mengurangi jumlah sampel kelas mayoritas agar distribusi data lebih seimbang. Pemilihan fitur dilakukan untuk memilih fitur paling relevan guna meningkatkan akurasi model dan mengurangi kompleksitas perhitungan, dengan teknik seperti korelasi Pearson, pohon keputusan, dan Recursive Feature Elimination (RFE). SMOTE juga membantu dalam proses ini dengan mengurangi bias fitur dominan pada kelas mayoritas, sehingga meningkatkan representasi kelas minoritas.

Pada tahap awal, model dikembangkan menggunakan Random Forest dengan TF-IDF tanpa menerapkan teknik balancing data. Namun, pendekatan ini sering kali menghadapi masalah ketidakseimbangan kelas, di mana model cenderung lebih akurat dalam memprediksi kelas mayoritas tetapi kurang mampu mengenali kelas

p-ISSN: 2775-4227
e-ISSN: 2775-4219

minoritas. Hal ini sesuai dengan temuan dalam penelitian sebelumnya yang menunjukkan bahwa model pembelajaran mesin cenderung bias terhadap kelas yang memiliki jumlah sampel lebih banyak, sehingga mengakibatkan rendahnya performa untuk kelas dengan jumlah data yang lebih sedikit [17]–[19]. Untuk mengatasi ketidakseimbangan ini, diterapkan Synthetic Minority Over-sampling Technique (SMOTE), yang menghasilkan sampel sintetis untuk meningkatkan representasi kelas minoritas. Metode ini telah terbukti efektif dalam meningkatkan performa model, terutama dalam meningkatkan recall, karena model memperoleh lebih banyak variasi sampel dari kelas yang kurang terwakili [20]. Namun, salah satu kelemahan SMOTE adalah potensi overfitting terhadap data sintetis, yang dapat menyebabkan penurunan presisi dan berkurangnya generalisasi model terhadap data baru [21]. Untuk mengurangi dampak overfitting, SMOTE dikombinasikan dengan Random Under-Sampling (RUS), yang bertujuan untuk mengurangi jumlah sampel dari kelas mayoritas sehingga distribusi data menjadi lebih seimbang. Beberapa penelitian sebelumnya telah menunjukkan bahwa kombinasi SMOTE dan RUS dapat meningkatkan keseimbangan antara presisi dan recall, karena pendekatan ini tidak hanya memperbanyak kelas minoritas tetapi juga mengurangi dominasi kelas mayoritas dalam pelatihan model [19]. Dengan demikian, strategi ini diharapkan dapat meningkatkan performa model secara keseluruhan dalam menangani ketidakseimbangan data pada klasifikasi teks.

Selain teknik balancing data, penelitian ini juga menerapkan GridSearchCV untuk mencari kombinasi hyperparameter terbaik dalam algoritma RF. Optimasi parameter ini mencakup jumlah pohon dalam hutan (n_estimators), kedalaman maksimum pohon (max_depth), dan jumlah minimum sampel yang diperlukan untuk membagi node (min_samples_split), yang bertujuan untuk meningkatkan stabilitas dan akurasi model [22]. Evaluasi model tidak hanya dilakukan berdasarkan akurasi, tetapi juga dengan F1-score dan Confusion Matrix untuk memahami bagaimana model menangani kesalahan klasifikasi [23].

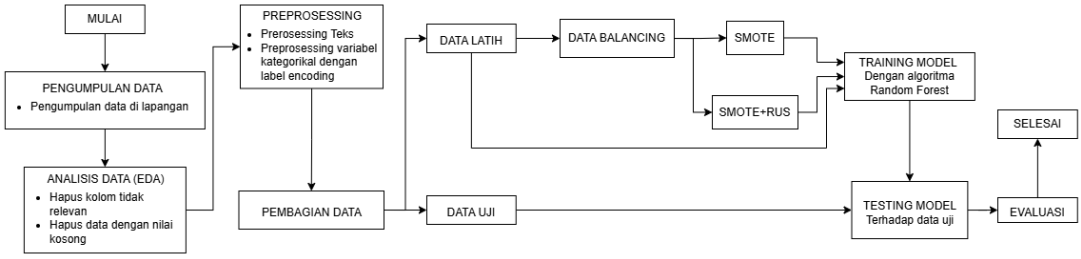
Hasil penelitian ini dibandingkan dengan temuan penelitian sebelumnya yang menerapkan pendekatan yang berbeda, seperti pada penelitian yang berjudul "Implementasi Algoritma Naïve Bayes untuk Memprediksi Kerusakan Sepeda Motor", penelitian ini mengembangkan model prediksi kerusakan sepeda motor menggunakan algoritma Naïve Bayes. Hasilnya menunjukkan bahwa metode ini dapat memberikan prediksi yang cukup akurat berdasarkan data historis perbaikan [1]; Peneltian terdahulu tentang "Klasifikasi Kerusakan Mesin Sepeda Motor Menggunakan Metode Neural Network Backpropagation", Penelitian ini merancang model machine learning menggunakan metode Neural Network Backpropagation untuk memprediksi jenis kerusakan mesin sepeda motor. Hasil penelitian menunjukkan bahwa model ini mampu mengklasifikasikan jenis kerusakan dengan tingkat akurasi yang tinggi [24]; Penelitian terdahulu berjusul "Penerapan Metode K-Nearest Neighbors (KNN) pada Bearing", penelitian ini menerapkan metode k-Nearest Neighbors untuk memprediksi Root Mean Square (RMS) bearing, yang merupakan indikator kondisi mesin. Meskipun tidak secara langsung terkait dengan sepeda motor, pendekatan ini menunjukkan potensi penggunaan kNN dalam prediksi kerusakan mesin [25]; dan pada penelitian terdahulu berjudul "Machine Learning untuk Prediksi Kegagalan Mesin dalam Sistem Perawatan Prediktif", penelitian ini memanfaatkan machine learning untuk mengklasifikasikan kegagalan mesin dalam membangun sistem perawatan prediktif. Hasilnya menunjukkan bahwa pendekatan ini dapat meningkatkan efisiensi dan akurasi dalam mendeteksi potensi kegagalan mesin [26].

Berdasarkan tantangan yang telah diidentifikasi dalam klasifikasi kerusakan sepeda motor, penelitian ini bertujuan untuk mengevaluasi kinerja algoritma Random Forest dengan berbagai teknik pemrosesan data dan balancing data, seperti TF-IDF, SMOTE, dan RUS. Dengan adanya kombinasi teknik ini, diharapkan model yang dikembangkan dapat menghasilkan prediksi yang lebih akurat, mengurangi bias terhadap kelas mayoritas, serta meningkatkan kemampuan dalam mengenali jenis kerusakan yang lebih jarang terjadi. Selain itu, optimasi hyperparameter menggunakan GridSearchCV dilakukan untuk meningkatkan performa model. Penelitian ini tidak hanya berkontribusi dalam pengembangan metode klasifikasi yang lebih baik, tetapi juga memberikan wawasan mengenai efektivitas teknik balancing data dalam meningkatkan performa model pembelajaran mesin dalam domain otomotif. Penelitian ini bertujuan untuk mengevaluasi kinerja algoritma *Random Forest* dalam klasifikasi kerusakan sepeda motor dengan menerapkan teknik praproses data (TF-IDF dan Label Encoding), metode penyeimbangan data (SMOTE dan RUS), serta optimasi *hyperparameter* menggunakan *GridSearchCV*, guna menghasilkan model diagnosis otomatis yang akurat, seimbang, dan dapat diandalkan dalam konteks industri perawatan kendaraan bermotor.

2. METODE PENELITIAN

Metodologi penelitian ini dirancang untuk mengevaluasi kinerja algoritma Random Forest dalam klasifikasi kerusakan motor berdasarkan data dari Bengkel PLAVIX. Tahapan penelitian terdiri dari beberapa langkah yang dapat dilihat pada Gambar 1 berikut.

p-ISSN: 2775-4227
e-ISSN: 2775-4219



Gambar 1. Diagram Alir Penelitian

2.1. Pengumpulan Data

Dalam penelitian ini, data diambil dari Bengkel PLAVIX yang bersifat privat. Dtaset berupa XLS kumpulan data servis yang terdiri dari 110 sampel data yang mencakup berbagai fitur terkait kerusakan sepeda motor. Fitur yang digunakan dalam penelitian ini meliputi Id, Merek Motor, Model Motor, Tipe Mesin, Transmisi, Jenis Motor, Gejala Kerusakan, dan Jenis Kerusakan. Data ini kemudian digunakan sebagai dasar dalam proses klasifikasi jenis kerusakan menggunakan model machine learning.

2.2. Analisis Data

Dataset yang digunakan dalam penelitian ini diperoleh dari Bengkel PLAVIX, yang berisi informasi terkait merek motor, model, jenis motor, gejala, dan jenis kerusakan. Pembacaan dataset dalam format Excel menggunakan Pandas, menghapus kolom yang tidak diperlukan seperti ID, Tipe Mesin, dan Transmisi karena tidak memberikan kontribusi dalam analisis klasifikasi serta menangani missing values pada kolom Gejala, Merek Motor, Model, Jenis Motor, dan Jenis Kerusakan. Dalam penelitian ini, variabel jenis kerusakan digunakan sebagai variabel target karena penelitian bertujuan untuk mengklasifikasikan jenis kerusakan berdasarkan data yang tersedia. Variabel fitur yang digunakan meliputi:

1. Gejala: Gejala kerusakan yang diuraikan dalam bentuk teks. Variabel ini diproses menggunakan transformasi TF-IDF (Term Frequency-Inverse Document Frequency) untuk mengubah teks menjadi representasi numerik.
2. Merek Motor: Digunakan untuk mengevaluasi apakah ada hubungan antara merek dan jenis kerusakan yang sering terjadi.
3. Model Motor: Beberapa model memiliki karakteristik berbeda dalam komponen dan daya tahan, sehingga dapat mempengaruhi pola kerusakan.
4. Jenis Motor: Dikategorikan menjadi bebek dan sport untuk membantu dalam pengelompokan data kerusakan.

2.3. Prepocessing Data

Tujuan tahap ini adalah membersihkan data agar sesuai dengan esensi setiap kata agar mudah diidentifikasi [27], [28]. Setelah data terkumpul, tahap selanjutnya adalah preprocessing yang bertujuan untuk membersihkan dan menyiapkan data agar siap untuk dianalisis lebih lanjut. Proses ini penting untuk memastikan bahwa data yang digunakan tidak mengandung noise atau elemen yang dapat mengganggu keakuratan model. Proses preprocessing terdiri dari beberapa tahapan yang harus dilalui agar diperoleh data yang bersih dan dapat dievaluasi.

a. Casefolding

Pada preprocessing tahap pertama dilakukan casefolding yaitu seluruh teks diubah menjadi huruf kecil untuk menghindari perbedaan penafsiran antara huruf besar dan huruf kecil. Proses ini juga mencakup penghapusan karakter yang tidak relevan seperti angka, simbol, atau tanda baca yang tidak mendukung analisis sentimen. Selain itu, pada tahap ini baris baru yang tidak perlu dibersihkan dan spasi ganda dihilangkan, yang dapat menyebabkan analisis tidak akurat [29].

b. Tokenization

Langkah penting dalam analisis leksikal adalah tokenisasi, yang bertujuan untuk memecah teks atau kalimat menjadi bagian kecil yang disebut sebagai "token", yang dapat berupa kata, frasa, atau karakter tergantung pada tingkat granularitas yang diinginkan [30].

c. Normalisasi Teks

Tahap selanjutnya adalah normalisasi teks, yang bertujuan untuk mengubah kata-kata dengan terminologi yang bervariasi atau tidak konsisten menjadi bentuk yang lebih umum atau standar. Hal ini penting agar model dapat lebih memahami kata-kata dan mengklasifikasikannya dengan benar. Misalnya, bahasa gaul, singkatan,

p-ISSN: 2775-4227
e-ISSN: 2775-4219

atau kata-kata dengan berbagai variasi ejaan akan diubah menjadi bentuk yang lebih umum agar lebih mudah dikenali dan diproses lebih lanjut [31].

d. Remove Stopwords

Pada tahap ini, kata-kata yang terdapat dalam daftar kamus stopword dihilangkan. Kata-kata berhenti adalah kata-kata yang sering muncul dalam teks tetapi tidak memberikan kontribusi signifikan terhadap analisis sentimen, seperti kata sambung, artikel, atau kata-kata umum lainnya. Tujuan menghilangkan kata-kata berhenti adalah untuk mengurangi kompleksitas data dan memungkinkan model untuk fokus pada kata-kata yang lebih penting dan relevan dalam mengidentifikasi sentimen yang ada dalam teks [32].

e. Stemming

Peran tahap stemming adalah mengubah kata turunan menjadi bentuk kata dasarnya. Proses ini penting agar model dapat mengenali kata-kata yang mirip meskipun bentuknya berubah, misalnya kata "running" akan diubah menjadi "run". Dengan menggunakan stemming, model dapat mengidentifikasi akar kata dengan lebih konsisten dan meningkatkan akurasi analisis sentimen yang dilakukan [33].

f. Encoding Fitur Kategorikal

Dalam penelitian ini, dilakukan proses encoding terhadap fitur kategorikal menggunakan Label Encoding dengan bantuan LabelEncoder dari scikit-learn. Fitur kategorikal yang diencoding meliputi "Merek Motor", "Model", dan "Jenis Motor", sedangkan variabel target yang diencoding adalah "Jenis Kerusakan". Proses encoding diawali dengan membangun objek LabelEncoder untuk setiap kolom kategorikal, kemudian setiap nilai dalam kolom tersebut dikonversi ke bentuk numerik sesuai dengan label yang dihasilkan oleh encoder.

g. Vectorization (TF-IDF)

Dalam penelitian ini, fitur Gejala, yang berupa teks deskripsi gejala kerusakan motor, diubah menjadi representasi numerik menggunakan TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF adalah teknik pengolahan teks yang memberikan bobot pada setiap kata berdasarkan pentingnya dalam sebuah dokumen dan dalam keseluruhan dataset. Formula dasar TF-IDF adalah:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t) \quad (1)$$

- Term Frequency (TF): TF (Term Frequency): Mengukur frekuensi relatif kemunculan kata t dalam dokumen d.
- Inverse Document Frequency (IDF): Mengukur seberapa unik kata t di seluruh dokumen dalam dataset.

Untuk memproses data teks menjadi numerik, TF-IDF digunakan sebagai metode utama untuk mengubah teks deskripsi gejala menjadi representasi numerik yang bermakna, sehingga dapat diproses oleh algoritma machine learning seperti Random Forest. TF-IDF memberikan bobot pada kata berdasarkan pentingnya dalam dokumen, yang menjadikannya ideal untuk menangani data teks dengan struktur sederhana dan sering digunakan dalam klasifikasi teks karena kemampuannya meningkatkan pemahaman algoritma terhadap konteks dan hubungan antar kata [34]. Dengan komponen IDF, kata-kata umum yang sering muncul, seperti "dan", "di", atau "ke", diberi bobot lebih kecil, sehingga menonjolkan kata-kata yang lebih relevan terhadap konteks gejala kerusakan [35]. Selain itu, konfigurasi ngram_range=(1, 2) memungkinkan TF-IDF menangkap tidak hanya kata tunggal (unigram) tetapi juga pasangan kata (bigram), sehingga mampu merepresentasikan pola hubungan antar kata dengan lebih baik dalam deskripsi gejala [36]. TF-IDF juga lebih efisien dibandingkan metode berbasis embedding seperti Word2Vec atau GloVe untuk dataset kecil dan terbukti meningkatkan akurasi pada tugas klasifikasi teks [37].

2.4. Pembagian Data

Dalam penelitian ini, pembagian data 80:20 dipilih dengan mempertimbangkan beberapa faktor penting yang mendukung kualitas pelatihan dan evaluasi model, terutama untuk dataset kecil seperti 110 data. Berikut adalah penjelasannya:

1. Memaksimalkan Data Latih

Dengan menggunakan 80% data untuk pelatihan, model memiliki lebih banyak data untuk mempelajari pola dan hubungan antar fitur. Hal ini sangat penting, terutama untuk dataset yang relatif kecil, agar model dapat memahami pola secara mendalam [38].

2. Menjaga Generalisasi Model

Proporsi 20% data uji dianggap cukup untuk mengevaluasi performa model pada data yang belum pernah dilihat sebelumnya. Rasio ini memastikan bahwa model tidak hanya bekerja dengan baik pada data pelatihan tetapi juga memiliki kemampuan generalisasi yang baik terhadap data baru [39].

3. Rekomendasi Praktik Standar

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Pembagian 80:20 merupakan praktik umum yang sering digunakan dalam penelitian machine learning. Rasio ini memberikan keseimbangan antara pelatihan yang memadai dan evaluasi yang representatif, terutama untuk dataset kecil [40].

4. Ukuran Dataset yang Relatif Kecil

Untuk dataset kecil seperti 110 data, pembagian ini memberikan jumlah data latih yang cukup besar untuk pelatihan model tanpa mengorbankan kualitas evaluasi pada data uji [41].

Dengan demikian, pembagian data 80:20 dalam penelitian ini dipilih untuk memastikan model memiliki data pelatihan yang cukup, performa evaluasi yang representatif, dan generalisasi yang baik.

2.5. Balancing Data

Dalam penelitian ini, dilakukan penyeimbangan data (data balancing) untuk mengatasi ketimpangan distribusi kelas (class imbalance) pada dataset. Ketidakseimbangan data dapat menyebabkan model cenderung lebih akurat dalam memprediksi kelas mayoritas, sementara performa pada kelas minoritas menjadi buruk karena model tidak mendapatkan cukup representasi dari kelas tersebut [35]. Oleh karena itu, digunakan kombinasi teknik oversampling dan undersampling untuk meningkatkan kualitas data latih.

Teknik yang digunakan dalam penelitian ini adalah:

1. SMOTE (Synthetic Minority Over-sampling Technique), yaitu metode untuk menghasilkan data tambahan dari kelas minoritas, mengatasi kekurangan datanya, dengan menghasilkan contoh pada garis yang menghubungkan suatu titik dan K tetangga terdekatnya [41].
2. Random UnderSampling (RUS), yaitu metode pengurangan sampel kelas mayoritas secara acak untuk menyeimbangkan distribusi kelas. RUS dipilih sebagai pelengkap SMOTE karena dapat mengurangi dominasi kelas mayoritas dalam dataset, meskipun berisiko kehilangan informasi yang berharga [31].
3. Kombinasi SMOTE dan RUS digunakan untuk mendapatkan keseimbangan optimal, di mana SMOTE menambah variasi pada kelas minoritas, sedangkan RUS mencegah model menjadi terlalu bias terhadap data sintetis yang dihasilkan oleh SMOTE [31].

Setelah proses balancing, dilakukan visualisasi distribusi kelas menggunakan countplot untuk melihat perbedaan proporsi kelas sebelum dan sesudah balancing. Hal ini penting untuk memastikan bahwa teknik balancing yang diterapkan berhasil dalam mengurangi ketimpangan kelas tanpa mengubah karakteristik utama dataset. Untuk mendapatkan kinerja model yang optimal, dilakukan proses penyetelan hiperparameter (hyperparameter tuning) menggunakan teknik Grid Search dengan validasi silang Stratified K-Fold Cross Validation (5-fold) [27]. Grid Search dipilih karena mampu menguji berbagai kombinasi parameter secara sistematis untuk menemukan kombinasi terbaik yang menghasilkan performa model tertinggi [34]. Parameter yang diuji dalam penelitian ini meliputi:

- n_estimators (jumlah pohon dalam Random Forest): 50, 100, 200
- max_depth (kedalaman maksimum pohon keputusan): 10, 15, 20
- min_samples_split (jumlah minimum sampel yang dibutuhkan untuk melakukan pemisahan node): 2, 5, 10

dari parameter default yang digunakan untuk teknik grid search dalam mencari kombinasi terbaik yang menghasilkan performa yang bagus dengan nilai *max_depth*:20, *min_samples_split*:2, *n_estimator*:200 serta mendapatkan best score: 0,84.Stratified K-Fold digunakan untuk memastikan bahwa proporsi kelas pada setiap fold tetap seimbang, sehingga hasil evaluasi lebih representatif dan mengurangi risiko bias akibat ketidakseimbangan kelas [28]. Pemilihan model terbaik dilakukan berdasarkan nilai F1-Score tertinggi untuk mengatasi ketidakseimbangan kelas dalam dataset [42].

2.6. Training, Testing, dan Evaluation Model

Setelah mendapatkan model terbaik dari proses hyperparameter tuning, dilakukan pelatihan dan evaluasi model pada dataset dengan Random Forest sebagai baseline model. Evaluasi dilakukan dengan membandingkan tiga skenario:

a. Baseline Model: Model yang dilatih tanpa teknik balancing data.
b. SMOTE Model: Model yang dilatih dengan data hasil oversampling menggunakan SMOTE.
c. SMOTE + RUS Model: Model yang dilatih dengan data hasil kombinasi SMOTE dan Random Undersampling.

Kinerja model diukur menggunakan metrik evaluasi berikut:

a. Akurasi:

Akurasi mengukur persentase prediksi benar terhadap total data uji. Rumus akurasi adalah:

p-ISSN: 2775-4227
e-ISSN: 2775-4219

$$Akurasi = \frac{\text{Jumlah Prediksi Benar}}{\text{Total Data Uji}} \quad (2)$$

Referensi: [43].

b. Presisi (Precision): Mengukur seberapa banyak prediksi positif yang benar [44]
c. Recall: Mengukur seberapa banyak data aktual positif yang berhasil diprediksi dengan benar .
d. F1-Score: Harmonik rata-rata precision dan recall [44].
e. Rumus:

$$Precision = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Positive (FP)}} \quad (3)$$

$$Recall = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Negative (FN)}} \quad (4)$$

$$F1\ Score = 2 \cdot \frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recall}} \quad (5)$$

- True Positive (TP): Data yang benar-benar positif dan diprediksi positif.
- False Positive (FP): Data yang sebenarnya negatif tetapi diprediksi positif.
- False Negative (FN): Data yang sebenarnya positif tetapi diprediksi negatif

f. Waktu Pelatihan: Mengukur efisiensi komputasi dari masing-masing skenario model.

Hasil evaluasi model divisualisasikan dalam bentuk diagram batang untuk membandingkan skor akurasi, presisi, recall, F1-score, dan waktu pelatihan dari ketiga skenario. Selain itu, dilakukan visualisasi Confusion Matrix untuk melihat distribusi kesalahan prediksi pada setiap model. Matriks kebingungan adalah tabel yang menunjukkan distribusi prediksi model terhadap kelas sebenarnya. Elemen-elemen matriks mencakup True Positive (TP), False Positive (FP), True Negative (TN), dan False Negative (FN) [44]. Rumus umum elemen matriks:

$$Confusion\ Matrix = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \quad (6)$$

Dengan tahapan-tahapan di atas, penelitian ini bertujuan untuk menghasilkan model Random Forest yang optimal dalam mengklasifikasikan jenis kerusakan motor berdasarkan data dari Bengkel PLAVIX. Hasil dari penelitian ini diharapkan dapat membantu bengkel dalam mendiagnosis kerusakan motor secara lebih akurat dan efisien.

3. HASIL DAN PEMBAHASAN

3.1. Hasil Penelitian

3.1.1. *Pengumpulan Data*

Dataset yang digunakan dalam penelitian ini terdiri dari 110 data sampel yang merepresentasikan berbagai jenis kerusakan pada motor manual dengan mesin karburator 4-tak. Dataset ini memiliki 8 variabel yang dikumpulkan secara manual di Bengkel PLAVIX, Karanggeneng, Semarang, dengan sempel 3 data pertama dapat dilihat pada Tabel 1 berikut ini:

Tabel 2. Sempel dataset awal

ID	Merek Motor	Model	Tipe Mesin	Transmisi	Jenis Motor	Jenis Kerusakan	Gejala
1	Honda	Astrea Prima	4 Tak	Manual	Bebek	Bering Roda	Motor terasa tidak stabil
2	Suzuki	Satria F	4 Tak	Manual	Sport	Shock Depan	Timbulnya getaran pada saat motor berjalan
3	Hinda	Megapro	4 Tak	Manual	Sport	Rantai Keteng	Mesin motor terasa bergetar

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Dari 8 variabel tersebut dipilih untuk menjadi variabel target dan variabel fitur. Untuk varibael targetnya adalah ’Jenis Kerusakan’ dan untuk variabel fiturnya ada ’Merek Motor’,’Model’,’Jenis Motor’ dan ’Gejala’. Untuk variabel yang dipilih karena di anggap lebih relevan dan efektif karena variabel yang tidak dipilih yaitu ’Id’, ’Tipe Mesin’ dan ’Transmisi’ hanya digunakan untuk penomoran, serta nilainya seragam (konstan) dalam dataset dan tidak memberikan kontribusi informasi tambahan pada proses klasifikasi.

3.1.2. Analisis Data

Tabel 3. Sampel dataset setetelah dibersihkan

ID	Merek Motor	Model	Jenis Motor	Jenis Kerusakan	Gejala
1	Honda	Astrea Prima	Bebek	Bering Roda	Motor terasa tidak stabil
2	Suzuki	Satria F	Sport	Shock Depan	Timbulnya getaran pada saat motor berjalan
3	Hinda	Megapro	Sport	Rantai Keteng	Mesin motor terasa bergetar

3.1.3. *Preprocessing* Teks

Di proses ini dataset yang diperoleh dilakukan preprocessing dengan beberapa tahapan secara berurutan untuk memproses teks ini menjadi fitur numerik serta mempersiapkan dataset agar dapat digunakan. Untuk sempel dataset original sebelum dilakukan preprocessing dapat dilihat pada Tabel 3.

Tabel 4. Sampel dataset variable teks original

Original Teks
Motor terasa tidak stabil
Motor kehilangan tenaga atau tenaganya berkurang
Kondisi noken as masih bagus tetapi kepala mesin/cylinder head masih ada suara berisik

Tahap-tahap pemrosesan yang dilakukan secara berurutan sebagai berikut:

a. Casefolding

Tabel 5. Hasil sempel dataset teks setelah case folding dan punctuation removal

Casefolding
motor terasa tidak stabil
motor kehilangan tenaga atau tenaganya berkurang
kondisi noken as masih bagus tetapi kepala mesin cylinder head masih ada suara berisik

b. Tokenization

Tabel 6. Hasil sempel dataset teks setelah tokenization

Tokenization
['motor', 'terasa', 'tidak', 'stabil']
['motor', 'kehilangan', 'tenaga', 'atau', 'tenaganya', 'berkurang']
['kondisi', 'noken', 'as', 'masih', 'bagus', 'tetapi', 'kepala', 'mesin', 'cylender', 'head', 'masih', 'ada', 'suara', 'berisik']

c. Normalization

Tabel 7. Hasil sempel dataset teks setelah normalization

Normalization
['motor', 'terasa', 'tidak', 'stabil']
['motor', 'kehilangan', 'tenaga', 'atau', 'tenaganya', 'berkurang']
['kondisi', 'noken', 'as', 'masih', 'bagus', 'tetapi', 'kepala', 'mesin', 'cylender', 'head', 'masih', 'ada', 'suara', 'berisik']

d. Stopword Removal

Tahap berikutnya adalah stopword removal, yaitu menghapus kata-kata umum yang tidak memiliki makna signifikan dalam analisis teks, seperti "yang", "dan", atau "di". Penghapusan stopword membantu mengurangi

p-ISSN: 2775-4227
e-ISSN: 2775-4219

kata-kata yang tidak memberikan kontribusi dalam proses TF-IDF, sehingga hanya kata-kata penting yang dipertimbangkan. Untuk hasil sempelnya dapat dilihat pada Tabel 7.

Tabel 8. Hasil sempel dataset teks setelah stopword removal

Stopword Removal
['motor', 'tidak', 'stabil']
['motor', 'kehilangan', 'tenaga', 'tenaganya', 'berkurang']
['kondisi', 'noken', 'as', 'bagus', 'kepala', 'mesin', 'cylender', 'head', 'berisik']

e. Stemming

Stemming untuk mengubah kata-kata ke bentuk dasarnya. Pada tahap ini, kata berimbuhan seperti "berlari" akan diubah menjadi "lari" agar kata dengan akar yang sama tidak dianggap berbeda. Dengan stemming, jumlah kata unik dalam teks berkurang, sehingga analisis teks menjadi lebih efisien dan akurat. Untuk hasil sempelnya dapat dilihat pada Tabel 8.

Tabel 9. Hasil sempel dataset teks setelah stemming (akhir)

Stemming
motor tidak stabil
motor hilang tenaga tenaga kurang
kondisi noken bagus kepala mesin cylender head suara berisik

f. Encoding Fitur Categorigal

Hasil encoding ini menghasilkan dataset yang sepenuhnya berupa nilai numerik, seperti yang ditampilkan pada Tabel 9.

Tabel 10. Hasil sempel dataset setelah di encoding

Merek Motor	#	Model	#	Jenis Motor	#	Jenis Kerusakan	#
Honda	0	Astrea prima	1	Bebek	0	Bering Roda	1
Suzuki	2	Satria F	10	Sport	1	Shock Depan	19
Honda	0	Megapro	9	Sport	1	Rantai Keteng	16
Honda	0	GL Pro	4	Sport	1	Klep	12
Suzuki	2	Shogun	11	Bebek	0	Bos Klep	3

g. TF-IDF

Tabel 11. Hasil sempel dataset teks setelah di tf-idf

”motor tidak stabil”	
motor	0.332336
motor tidak	0.455493
stabil	0.539524
tidak	0.316080
tidak stabil	0.539524

3.1.4. *Splitting Data*

3.1.5. *Data Balancing*

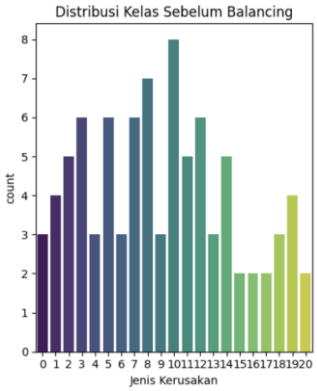
Balancing dataset penting dilakukan dalam model klasifikasi untuk mengatasi ketidakseimbangan antara kelas mayoritas dan minoritas [45], [46]. Ketidakseimbangan ini dapat menyebabkan model cenderung memprediksi kelas mayoritas dengan akurasi tinggi, namun gagal mengenali kelas minoritas yang jarang terjadi, meskipun kelas tersebut mungkin lebih penting dalam konteks aplikasi tertentu [47]. Hal ini mengarah pada performa buruk pada kelas minoritas, meskipun model secara keseluruhan menunjukkan akurasi yang baik [48], [49].

Untuk mengatasi masalah ini, teknik seperti SMOTE (*Synthetic Minority Over-sampling Technique)* dan *Random Under Sampling* (RUS) digunakan untuk menyeimbangkan distribusi kelas dengan menambah data sintetis pada kelas minoritas atau mengurangi sampel dari kelas mayoritas. Dengan balancing, model dapat lebih

p-ISSN: 2775-4227
e-ISSN: 2775-4219

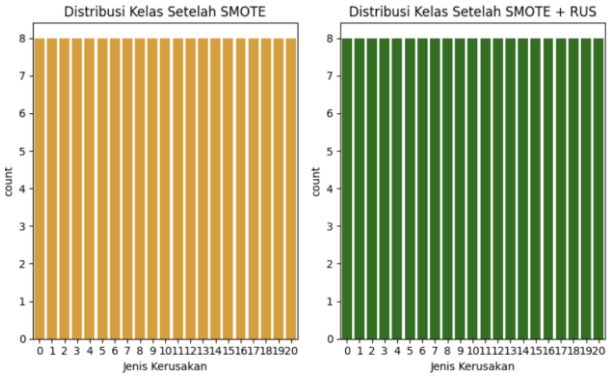
adil dalam memprediksi kedua kelas, meningkatkan akurasi dan kemampuan model dalam mengenali pola pada kelas minoritas, yang sering kali lebih kritikal untuk aplikasi dunia nyata.

Dalam proses ini ketidakseimbangan kelas dalam dataset dapat menyebabkan model cenderung lebih akurat dalam memprediksi kelas mayoritas, sementara performa pada kelas minoritas menjadi buruk. Distribusi kelasnya dapat dilihat pada Gambar 2.



Gambar 2. Distribusi kelas sebelum balancing

Untuk mengatasi hal ini, kombinasi SMOTE dan RUS diterapkan untuk menjaga keseimbangan data dan hasilnya dapat dilihat pada Gambar 3.



Gambar 3. Distribusi kelas setelah balancing SMOTE dan SMOTE+RUS

Setelah balancing data, distribusi kelas menjadi lebih merata, yang diharapkan berkontribusi pada peningkatan performa model dalam mengenali pola kerusakan motor yang lebih jarang terjadi.

3.1.6. *Training, Testing,* dan *Evaluation Model*

Pada penelitian ini, dilakukan evaluasi terhadap tiga model klasifikasi menggunakan Random Forest dengan berbagai metode penyeimbangan data, yaitu model baseline tanpa balancing, model dengan Synthetic Minority Over-sampling Technique (SMOTE), serta model dengan kombinasi SMOTE dan Random Under Sampling (RUS). Kinerja model dibandingkan berdasarkan metrik akurasi, presisi, recall, F1-score, serta waktu training. Hasil evaluasi ditampilkan pada Tabel 11 berikut.

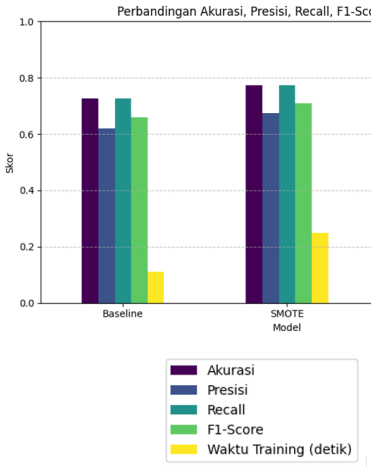
Tabel 12. Hasil pengujian

Model	Akurasi	Presisi	Recall	F1-Score	Waktu (s)
Baseline	72.73%	62.12%	72.73%	66.06%	0.22
SMOTE	77.27%	67.42%	77.27%	70.91%	0.42
SMOTE+RUS	77.27%	64.39%	77.27%	69.09%	0.41

Berdasarkan hasil yang diperoleh, model dengan teknik SMOTE dan SMOTE + RUS menunjukkan peningkatan akurasi dibandingkan dengan model Baseline. Model Baseline memperoleh akurasi 72.73%, sementara model dengan SMOTE dan SMOTE + RUS meningkatkan akurasi menjadi 77.27%. Selain akurasi,

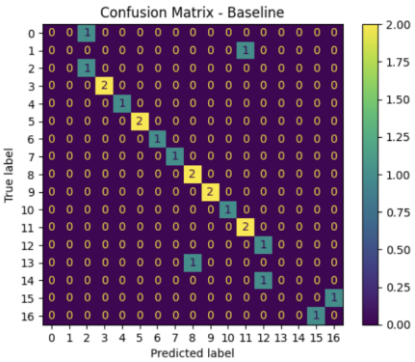
p-ISSN: 2775-4227
e-ISSN: 2775-4219

nilai presisi, recall, dan F1-score juga mengalami peningkatan setelah dilakukan balancing. Model SMOTE menunjukkan presisi lebih tinggi dibandingkan SMOTE + RUS, yaitu 67.42% dibandingkan 64.39%, tetapi F1-score dari SMOTE + RUS lebih rendah dibandingkan SMOTE (69.09% vs. 70.91%). Hal ini menunjukkan bahwa SMOTE cenderung lebih stabil dalam meningkatkan performa model dibandingkan kombinasi SMOTE + RUS, meskipun perbedaannya tidak terlalu signifikan. Dari segi waktu training, model dengan SMOTE dan SMOTE + RUS memiliki waktu komputasi hampir dua kali lipat dibandingkan model baseline (0.42 detik dan 0.41 detik dibandingkan 0.22 detik). Hal ini menunjukkan bahwa penyeimbangan data membutuhkan waktu tambahan dalam proses pelatihan, tetapi tetap dalam rentang waktu yang dapat diterima. Untuk memvisualisasikan perbandingan kinerja ketiga model, ditampilkan grafik batang pada Gambar 4 berikut.

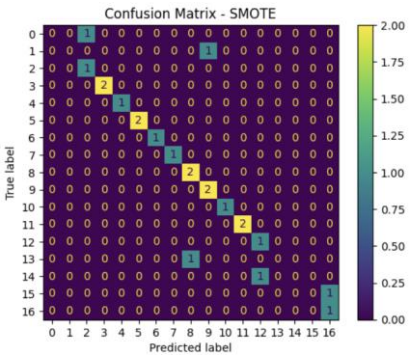


Gambar 4. Grafik perbandingan kinerja 3 model

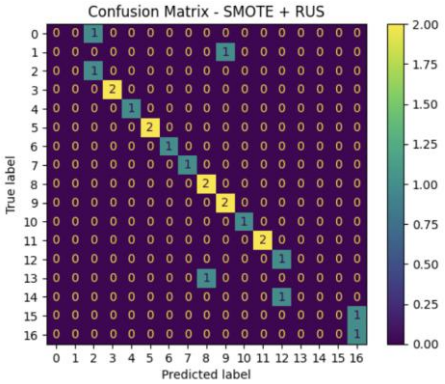
Selain itu, untuk melihat performa model dalam menangani klasifikasi setiap kelas, ditampilkan confusion matrix untuk masing-masing model pada Gambar 5, Gambar 6, dan Gambar 7.



Gambar 5. Confusion matrix baseline



Gambar 6. Confusion matrix SMOTE



Gambar 7. Confusion matrix SMOTE+RUS

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Dari confusion matrix, terlihat bahwa model dengan balancing menggunakan SMOTE dan SMOTE + RUS mampu mengurangi jumlah kesalahan klasifikasi dibandingkan model baseline, terutama pada kelas minoritas.

Kombinasi metode Random Forest (RF) dengan teknik Synthetic Minority Oversampling Technique (SMOTE) dan Random Under Sampling (RUS) efektif dalam menangani ketidakseimbangan data serta meningkatkan kinerja model klasifikasi [50]. Hal ini didukung oleh penelitian yang dilakukan oleh Istiqomah (2022) tentang Sistem Klasifikasi Ulasan Konsumen Menggunakan Algoritma Random Forest dan Synthetic Minority Oversampling Technique (SMOTE). Penelitian ini membahas penerapan algoritma Random Forest dan SMOTE untuk mengklasifikasikan ulasan konsumen ke dalam kategori positif dan negatif. Hasil penelitian menunjukkan bahwa penerapan SMOTE dan Random Under Sampling pada tahap pra-pemrosesan meningkatkan akurasi sekitar 3% dan skor AUC sekitar 4%, serta meningkatkan kinerja model dalam mengenali data yang awalnya dianggap minoritas [12]. Selain itu, Sari et al. dalam Jurnal Sistem dan Komputer mengkaji penggunaan RF dan SMOTE dalam klasifikasi ulasan konsumen dan menemukan bahwa metode ini mampu meningkatkan akurasi model secara signifikan. Penelitian lain oleh Dharmendra dalam Informatics for Educators and Professionals Journal juga menegaskan bahwa kombinasi SMOTE dan RUS dapat meningkatkan performa model klasifikasi pada data tidak seimbang.

3.2. Pembahasan

Hasil dari penelitian ini menunjukkan bahwa penggunaan metode SMOTE dan kombinasi SMOTE + RUS mampu meningkatkan kinerja model Random Forest dalam mengklasifikasikan jenis kerusakan motor. Model baseline tanpa teknik balancing hanya menghasilkan akurasi sebesar 72,73%, sedangkan dengan SMOTE meningkat menjadi 77,27% dan kombinasi SMOTE + RUS juga mencapai 77,27%. Walaupun akurasinya sama, terdapat perbedaan dari sisi presisi dan F1-score, di mana model SMOTE memiliki presisi sebesar 67,42% dan F1-score 70,91%, sedangkan SMOTE + RUS memiliki presisi 64,39% dan F1-score 69,09%.

Peningkatan ini menunjukkan bahwa balancing data memiliki peran penting dalam mengatasi ketimpangan distribusi kelas yang sebelumnya menyebabkan model lebih fokus pada prediksi kelas mayoritas. Secara analitis, metode SMOTE bekerja dengan cara membuat sampel sintetis dari kelas minoritas berdasarkan kedekatan antar sampel, sehingga model memperoleh variasi yang lebih banyak untuk kelas tersebut. Ini menjelaskan peningkatan recall yang signifikan (dari 72,73% menjadi 77,27%).

Namun, penggunaan SMOTE secara tunggal berisiko menghasilkan data sintetis yang terlalu mirip, sehingga berpotensi menyebabkan overfitting. Untuk mengatasi hal ini, kombinasi SMOTE + RUS digunakan, di mana RUS menghapus sebagian data dari kelas mayoritas secara acak. Pendekatan ini menghasilkan distribusi data yang lebih seimbang, mengurangi dominasi kelas mayoritas, dan meningkatkan kemampuan model dalam mengklasifikasikan kelas minoritas secara lebih adil. Meskipun presisi sedikit menurun, hal ini merupakan trade-off yang wajar untuk meningkatkan keseimbangan recall antar kelas.

Temuan ini konsisten dengan studi sebelumnya oleh Istiqamah & Rijal (2024), di mana akurasi model meningkat dari 73% menjadi 75% setelah menerapkan SMOTE + RUS. Penelitian oleh Aryanti dkk. (2023) juga menunjukkan peningkatan dari 86% menjadi 88% saat menggunakan RF + SMOTE pada klasifikasi berbasis teks. Hal ini membuktikan bahwa pendekatan kombinasi balancing data efektif tidak hanya pada dataset numerik, tetapi juga pada dataset berbasis teks seperti dalam penelitian ini yang menggunakan fitur gejala kerusakan sepeda motor dengan TF-IDF vectorization.

Selain itu, meskipun waktu pelatihan model dengan SMOTE dan SMOTE + RUS meningkat dari 0,22 detik (baseline) menjadi 0,42 dan 0,41 detik, peningkatan ini masih dalam batas yang sangat wajar dan tidak berdampak signifikan terhadap efisiensi sistem.

Secara keseluruhan, penerapan balancing data terbukti meningkatkan stabilitas dan performa klasifikasi dibandingkan hanya mengandalkan model baseline. Kombinasi TF-IDF, Label Encoding, serta SMOTE + RUS dalam model Random Forest menghasilkan model yang tidak hanya akurat, tetapi juga lebih adil dan adaptif terhadap distribusi kelas yang tidak seimbang.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa algoritma Random Forest yang dikombinasikan dengan teknik SMOTE dan RUS mampu mengatasi ketidakseimbangan data dan menghasilkan akurasi tertinggi sebesar 77,27%. Hasil ini lebih baik dibandingkan model baseline tanpa balancing data yang hanya mencapai akurasi 72,73%. Teknik TF-IDF terbukti efektif dalam merepresentasikan data teks gejala kerusakan, sehingga model dapat mengenali pola klasifikasi dengan lebih akurat.

Secara praktis, model ini dapat membantu bengkel dalam mendiagnosis kerusakan motor secara lebih cepat dan objektif, mengurangi ketergantungan pada pengalaman subjektif mekanik, serta meningkatkan efisiensi layanan.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Untuk pengembangan selanjutnya, disarankan agar penelitian ini diuji pada dataset yang lebih besar dan beragam, serta dikombinasikan dengan algoritma lain seperti XGBoost atau SVM. Penelitian lanjutan juga dapat mempertimbangkan fitur sensorik seperti suhu dan getaran untuk meningkatkan akurasi dan cakupan prediksi model.

DAFTAR PUSTAKA

[1] L. Epriliani, Mayadi, and R. W. P. Pamungkas, "Implementasi Algoritma Naïve Bayes Untuk Memprediksi Kerusakan Sepeda Motor Pada Bengkel Citra Djaya Motor," *J. Inform. Inf. Secur.*, vol. 3, no. 1, pp. 59–72, 2022, doi: 10.31599/jiforty.v3i1.1268.

[2] M. L. T. Alfianti and R. Supriyanto, "Perbandingan Kinerja Algoritma Random Forest, AdaBoost, dan XGBoost Dalam Memprediksi Resiko Penyakit Osteoporosis," *J. Ilmu Komput. dan Agri-Informatika*, vol. 11, no. 2, pp. 172–184, Nov. 2024, doi: 10.29244/jika.11.2.172-184.

[3] U. Sunarya and T. Haryanti, "Perbandingan Kinerja Algoritma Optimasi pada Metode Random Forest untuk Deteksi Kegagalan Jantung," *J. Rekayasa Elektr.*, vol. 18, no. 4, pp. 241–247, 2022, doi: 10.17529/jre.v18i4.26981.

[4] M. I. Arisani and M. Muljono, "Peningkatan Kinerja K-Nearest Neighbor menggunakan Bagging pada Permasalahan Ragam Kelas terhadap Pemeliharaan Prediktif Permesinan," *JUSTIN (Jurnal Sist. dan Teknol. Informasi)*, vol. 12, no. 2, pp. 373–379, 2024, doi: 10.26418/justin.v12i2.78503.

[5] M. A. Rayadin, M. Musaruddin, R. A. Saputra, and I. Isnawaty, "Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 2, pp. 111–119, 2024.

[6] A. Kurniawan, "Sistem Pakar Diagnosa Kerusakan Mesin Sepeda Motor dengan Menggunakan Metode Forward Chaining," *J. Ilmu Komputer, Tek. dan Multimed.*, vol. 1 No 2, no. 2, p. 446, 2023.

[7] B. E. S. Dewi, S. Haikal, H. . Sulistyowati, R. Fitriani, and D. Pranowo, "Penerapan Machine Learning Menggunakan Algoritma Random Forest untuk Prediksi Harga Mobil Bekas," *Jurbal Tridi*, vol. 2, no. 1, pp. 20–31, 2024.

[8] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Investigating class rarity in big data," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00301-0.

[9] S. K. Narayanasamy, Y. C. Hu, S. M. Qaisar, and K. Srinivasan, "Effective Preprocessing and Normalization Techniques for COVID-19 Twitter Streams with POS Tagging via Lightweight Hidden Markov Model," *J. Sensors*, vol. 2022, 2022, doi: 10.1155/2022/1222692.

[10] P. Cerda, G. Varoquaux, P. Cerda, and G. Varoquaux, "Encoding high-cardinality string categorical variables To cite this version : Encoding high-cardinality string categorical variables," 2020.

[11] E. Saputro and D. Rosiyadi, "Penerapan Metode Random Over-Under Sampling Pada Algoritma Klasifikasi Penentuan Penyakit Diabetes," *Bianglala Inform.*, vol. 10, no. 1, pp. 42–47, 2022, doi: 10.31294/bi.v10i1.11739.

[12] N. Istiqamah and M. Rijal, "Klasifikasi Ulasan Konsumen Menggunakan Random Forest dan SMOTE," *J. Syst. Comput. Eng.*, vol. 5, no. 1, pp. 66–77, 2024, doi: 10.61628/jsce.v5i1.1061.

[13] R. Aryanti, T. Misriati, and A. Sagiyanto, "Analisis Sentimen Aplikasi Primaku Menggunakan Algoritma Random Forest dan SMOTE untuk Mengatasi Ketidakseimbangan Data," *J. Comput. Syst. Informatics*, vol. 5, no. 1, pp. 218–227, 2023, doi: 10.47065/josyc.v5i1.4562.

[14] L. Dube and T. Verster, "Enhancing classification performance in imbalanced datasets: A comparative analysis of machine learning models," *Data Sci. Financ. Econ.*, vol. 3, no. 4, pp. 354–379, 2023, doi: 10.3934/DSFE.2023021.

[15] L. Xiang, "Application of an Improved TF-IDF Method in Literary Text Classification," *Adv. Multimed.*, vol. 2022, 2022, doi: 10.1155/2022/9285324.

[16] G. Dudek, "A Comprehensive Study of Random Forest for Short-Term Load Forecasting," *Energies*, vol. 15, no. 20, 2022, doi: 10.3390/en15207547.

[17] V. Rupapara, F. Rustam, H. F. Shahzad, A. Mehmood, I. Ashraf, and G. S. Choi, "Impact of SMOTE on Imbalanced Text Features for Toxic Comments Classification Using RVVC Model," *IEEE Access*, vol. 9, pp. 78621–78634, 2021, doi: 10.1109/ACCESS.2021.3083638.

[18] H. Hairani, A. Anggrawan, and D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.

p-ISSN: 2775-4227
e-ISSN: 2775-4219

[19] T. Fulazzaky, A. Saefuddin, and A. M. Soleh, "Evaluating Ensemble Learning Techniques for Class Imbalance in Machine Learning : A Comparative Analysis of Balanced Random," vol. 11, no. 4, pp. 969–980, 2024, doi: 10.15294/sji.v11i4.15937.

[20] A. A. G. W. S. Erlangga, I. G. A. Gunadi, and I. M. G. Sunarya, "Kombinasi Oversampling dan Undersampling dalam Menangani Class Imbalanced dan Overlapping pada Klasifikasi Data Bank Marketing," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 7, no. 1, pp. 32–42, 2024, doi: 10.31598/jurnalresistor.v7i1.1515.

[21] N. Suryana, P. Pratiwi, and R. T. Prasetio, "Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting," *IJCIT (Indonesian J. Comput. Inf. Technol.*, vol. 6, no. 1, pp. 31–37, May 2021, doi: 10.31294/ijcit.v6i1.9545.

[22] T. O. Omotehinwa and D. O. Oyewola, "Hyperparameter Optimization of Ensemble Models for Spam Email Detection," *Appl. Sci.*, vol. 13, no. 3, 2023, doi: 10.3390/app13031971.

[23] Jubeile Mark Baladjay, Nisce Riva, Ladine Ashley Santos, Dan Michael Cortez, Criselle Centeno, and Ariel Antwaun Rolando Sison, "Performance evaluation of random forest algorithm for automating classification of mathematics question items," *World J. Adv. Res. Rev.*, vol. 18, no. 2, pp. 034–043, 2023, doi: 10.30574/wjarr.2023.18.2.0762.

[24] M. H. Ibrahim, "Klasifikasi Kerusakan Mesin Sepeda Motor menggunakan Metode Neural Network Backpropagation," *J. GEEJ*, vol. 7, no. 2, 2024.

[25] Anggi Priliani Yulianto and S. Darwis, "Penerapan Metode K-Nearest Neighbors (kNN) pada Bearing," *J. Ris. Stat.*, vol. 1, no. 1, pp. 10–18, 2021, doi: 10.29313/jrs.v1i1.16.

[26] N. Hafidhoh, A. P. Atmaja, G. N. Syaifuddiin, I. B. Sumafta, S. M. Pratama, and H. N. Khasanah, "Machine Learning untuk Prediksi Kegagalan Mesin dalam Predictive Maintenance System," *J. Masy. Inform.*, vol. 15, no. 1, pp. 56–66, 2024, doi: 10.14710/jmasif.15.1.63641.

[27] S. Widodo, H. Brawijaya, and S. Samudi, "Stratified K-fold cross validation optimization on machine learning for prediction," *Sinkron*, vol. 7, no. 4, pp. 2407–2414, 2022, doi: 10.33395/sinkron.v7i4.11792.

[28] S. Maldonado, J. López, and A. Iturriaga, "Out-of-time cross-validation strategies for classification in the presence of dataset shift," *Appl. Intell.*, vol. 52, no. 5, pp. 5770–5783, 2022, doi: 10.1007/s10489-021-02735-2.

[29] C. Cahyaningtyas, Y. Nataliani, and I. R. Widiasari, "Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE," *Aiti*, vol. 18, no. 2, pp. 173–184, 2021, doi: 10.24246/aiti.v18i2.173-184.

[30] J. E. Br Sinulingga and H. C. K. Sitorus, "Analisis Sentimen Opini Masyarakat terhadap Film Horor Indonesia Menggunakan Metode SVM dan TF-IDF," *J. Manaj. Inform.*, vol. 14, no. 1, pp. 42–53, 2024, doi: 10.34010/jamika.v14i1.11946.

[31] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Inf.*, vol. 14, no. 1, 2023, doi: 10.3390/info14010054.

[32] M. Samantri and Afiyati, "Perbandingan Algoritma Support Vector Machine dan Random Forest untuk Analisis Sentimen Terhadap Kebijakan Pemerintah Indonesia Terkait Kenaikan Harga BBM Tahun 2022," *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 8, no. 1, pp. 1–9, 2024, doi: 10.35870/jtik.v8i1.1202.

[33] A. H. Sial, S. Yahya, and S. Rashdi, "Comparative Analysis of Data Visualization Libraries Matplotlib and Seaborn in Python," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 1, pp. 277–281, 2021, doi: 10.30534/ijatcse/2021/391012021.

[34] Sukamto, Hadiyanto, and Kurnianingsih, "KNN Optimization Using Grid Search Algorithm for Preeclampsia Imbalance Class," *E3S Web Conf.*, vol. 448, 2023, doi: 10.1051/e3sconf/202344802057.

[35] H. Xu and J. A. Prozzi, "Effect of Data Imbalance on the Performance of Pavement Deterioration Models," *Transp. Res. Rec.*, vol. 2677, no. 12, pp. 201–211, 2023, doi: 10.1177/03611981231167427.

[36] Venkata Mahesh Babu Batta, "Human Language Data Processing using NLTK," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 628–634, 2024, doi: 10.48175/ijarsct-17685.

[37] J. H. M. Daniel Jurafsky, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson, 2023.

[38] M. Alden, N. Anargya, W. Ghozi, and F. A. Rafrastara, "Optimizing IoV Attack Detection using Random Under Sampling Techniques," vol. 10, no. 1, pp. 11–19, 2025, doi: 10.30591/jpit.v10i1.8034.

[39] N. I. Yaman, A. R. Juwita, S. Arum, P. Lestari, and S. Faisal, "Perbandingan Kinerja Algoritma Decision

p-ISSN: 2775-4227
e-ISSN: 2775-4219

Tree dan Random Forest untuk Klasifikasi Nutrisi pada Makanan Cepat Saji," pp. 184–195, 2024, doi: 10.33364/algoritma/v.21-2.1649.

[40] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," *7th Int. Mardin Artuklu Sci. Res. Conf.*, no. February, pp. 496–504, 2022.

[41] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.

[42] C. Magnolia, A. Nurhopipah, and B. A. Kusuma, "Penanganan Imbalanced Dataset untuk Klasifikasi Komentar Program Kampus Merdeka Pada Aplikasi Twitter," *Edu Komputika J.*, vol. 9, no. 2, pp. 105–113, 2022, doi: 10.15294/edukomputika.v9i2.61854.

[43] K. Alanne and S. Sierla, "An overview of machine learning applications for smart buildings," *Sustain. Cities Soc.*, vol. 76, no. July 2021, p. 103445, 2022, doi: 10.1016/j.scs.2021.103445.

[44] M. Faaique, "Overview of Big Data Analytics in Modern Astronomy," *Int. J. Math. Stat. Comput. Sci.*, vol. 2, pp. 96–113, 2023, doi: 10.59543/ijmscs.v2i.8561.

[45] M. Hafiz and B. Prayoga, "Analisis Pemilihan Jurusan pada Calon Siswa SMK Negeri 4 Palembang Pada Faktor Penentu Pemilihan Jurusan Menggunakan Association Rule dan Random Forest Analysis of Major Selection for Prospective Students of SMK Negeri 4 Palembang on Determining Factors f," vol. 4, no. 12, pp. 537–547, 2024.

[46] R. Yunanto and U. Budiyanto, "Implementasi XGBoost dan SMOTE untuk Meningkatkan Deteksi Transaksi Fraud di Industri Jasa Keuangan Implementing Xgboost Models For Enhanced Detection Of Fraud Transaction In Financial Services Industries," vol. 4, no. 11, pp. 525–535, 2024.

[47] A. Mu *et al.*, "Optimasi Logistic Regression dan Random Forest untuk Deteksi Berita Hoax Berbasis Hyperparameter Optimization of Logistic Regression and Random Forest for Hoax News Detection Using TF-IDF Text Representation," vol. 4, no. 8, pp. 381–392, 2024.

[48] A. S. Asaury *et al.*, "Prediksi Jumlah Pasien Masuk Rumah Sakit Menggunakan Metode Random Forest PREDICTION OF THE NUMBER OF PATIENTS ADMITTED TO HOSPITAL USING," vol. 5, no. 2, pp. 447–459, 2025.

[49] D. Ariyana, E. D. Wahyuni, and N. Sembilu, "Perbandingan Kinerja Metode Binary Relevance , Classifier Chains , dan Label Powerset dalam Klasifikasi Multi-Label Data Pengaduan Sistem Informasi , Fakultas Ilmu Komputer , Universitas Pembangunan Nasional Veteran Jawa Comparison of Evaluation Results o," vol. 5, no. 3, pp. 615–623, 2025.

[50] U. Hasanah, A. mohamad Soleh, and K. Sadik, "Effect of Random Under sampling , Oversampling , and SMOTE on the Performance of Cardiovascular Disease Prediction Models terhadap Kinerja Model Prediksi Penyakit Kardiovaskular," *J. Mat. Stat. dan Komputasi*, vol. 21, no. 1, pp. 88–102, 2024, doi: 10.20956/j.v21i1.35552.