Perbandingan Logistic Regression dan Random Forest untuk Prediksi Respon Pelanggan Asuransi

Harliana*[1], Tito Prabowo[2], Ady Alzhava Nuary[3]
[1,2,3] Program Studi, Ilmu Komputer, Fakultas lmu Eksakta, Universitas Nahdlatul Ulama Blitar
*e-mail: *[1]harliana@unublitar.ac.id, [2]titoprabowo@unublitar.ac.id,* [3]adyalzhavanuary@gmail.com
*Correspondence author email: **

Abstrak
Industri asuransi kendaraan membutuhkan strategi pemasaran berbasis data untuk mengidentifikasi calon pelanggan yang berpotensi memberikan respon positif terhadap penawaran produk. Namun, prediksi respon pelanggan menghadapi tantangan berupa ketidakseimbangan kelas karena jumlah pelanggan yang tidak merespon jauh lebih besar dibandingkan pelanggan yang merespon. Penelitian ini bertujuan membandingkan performa algoritma Logistic Regression dan Random Forest dalam memprediksi respon pelanggan terhadap produk asuransi kendaraan menggunakan metode SMOTE. Dataset yang digunakan adalah Vehicle Insurance dari Kaggle. Hasil penelitian menunjukkan bahwa Random Forest memberikan performa terbaik dengan akurasi 0,80, F1-score kelas positif 0,59, dan ROC-AUC 0,88. Sementara itu, Logistic Regression memiliki *recall* kelas positif lebih tinggi sebesar 0,98, tetapi menghasilkan *precision* lebih rendah sebesar 0,35. Analisis *feature importance* menunjukkan bahwa *Previously_Insured*, *Vehicle_Damage*, dan *Age* merupakan faktor dominan yang memengaruhi respon pelanggan. Oleh karena itu, kombinasi Random Forest dan SMOTE dapat dipertimbangkan sebagai pendekatan yang efektif dalam memprediksi respon pelanggan terhadap produk asuransi kendaraan.
Kata kunci—Prediksi respon konsumen; Logistic Regression; Random Forest; *Machine learning*; *Data-driven marketing*

1. PENDAHULUAN

Industri asuransi khususnya asuransi kendaraan, mengalami perkembangan yang signifikan seiring dengan meningkatnya jumlah kendaraan dan kesadaran masyarakat terhadap manajemen risiko. Persaingan yang semakin ketat mendorong perusahaan asuransi untuk tidak hanya berfokus pada produk, tetapi juga pada pemahaman yang mendalam terhadap perilaku pelanggan dalam menentukan minat terhadap suatu layanan. Dalam konteks ini, kemampuan untuk mengidentifikasi pola dan kecenderungan pelanggan menjadi sangat penting guna meningkatkan efektivitas strategi pemasaran dan retensi pelanggan. Pertumbuhan jumlah data pelanggan memungkinkan perusahaan melakukan analisis yang lebih akurat untuk mendukung proses pengambilan keputusan [1], sehingga dapat membantu menentukan target pasar dan meningkatkan kinerja bisnis secara keseluruhan [2].

Namun, dalam praktiknya memprediksi respon pelanggan terhadap penawaran produk asuransi bukanlah hal yang sederhana. Tidak semua pelanggan yang menjadi target pemasaran akan memberikan respon positif, sehingga perusahaan sering menghadapi permasalahan dalam menentukan strategi promosi yang tepat sasaran. Akibatnya, kampanye pemasaran yang dilakukan berpotensi tidak efisien karena menyasar pelanggan yang tidak memiliki kecenderungan untuk membeli produk asuransi. Selain itu, permasalahan lain yang umum terjadi dalam data pelanggan adalah ketidakseimbangan distribusi kelas (*class imbalance*), di mana jumlah pelanggan yang tidak merespon jauh lebih besar dibandingkan dengan pelanggan yang merespon. Kondisi ini dapat menyebabkan model klasifikasi menjadi bias terhadap kelas mayoritas dan menurunkan kemampuan model dalam mendeteksi pelanggan yang berpotensi merespon, sehingga diperlukan pendekatan khusus untuk mengatasinya [3], [4].

Untuk mengatasi permasalahan tersebut, penelitian ini menggunakan pendekatan klasifikasi untuk mengelompokkan pelanggan ke dalam kategori respon positif atau negatif berdasarkan karakteristik yang dimiliki. Pendekatan ini memanfaatkan data historis pelanggan

untuk membangun model yang mampu mengenali pola serta memprediksi kecenderungan respon terhadap penawaran produk asuransi. Dari berbagai metode klasifikasi yang ada, Logistic Regression dan Random Forest termasuk algoritma yang sering diterapkan pada permasalahan klasifikasi biner [5] serta menunjukkan performa yang baik pada berbagai penelitian [6]. Logistic Regression memiliki model yang relatif sederhana sehingga hasil prediksinya lebih mudah diinterpretasikan [7], sedangkan Random Forest mampu mengenali hubungan yang lebih kompleks dan meningkatkan akurasi prediksi melalui kombinasi sejumlah pohon keputusan [8]. Oleh karena itu, kedua metode tersebut dipilih untuk memodelkan prediksi respon pelanggan pada penelitian ini.

Namun demikian, beberapa penelitian sebelumnya dalam konteks prediksi respon pelanggan masih menunjukkan keterbatasan. Sebagian studi cenderung hanya menggunakan satu algoritma klasifikasi seperti penelitian [9] dan [10] sehingga belum memberikan gambaran komparatif mengenai kinerja model yang berbeda dalam menangani karakteristik data yang kompleks. Selain itu, banyak penelitian yang belum mempertimbangkan permasalahan ketidakseimbangan data (*class imbalance*) [11], [12]. Padahal kondisi ini umum terjadi pada data pelanggan dan dapat menyebabkan model bias terhadap kelas mayoritas [13]. Keterbatasan lainnya terletak pada evaluasi performa model yang belum dilakukan secara menyeluruh, terutama dalam menilai keseimbangan antara kemampuan model mengenali kelas minoritas dan tingkat kesalahan prediksi [3].

Berdasarkan uraian tersebut, masih terdapat celah penelitian mengenai prediksi respon pelanggan asuransi, khususnya terkait terbatasnya penelitian yang membandingkan kinerja beberapa algoritma klasifikasi pada data pelanggan yang tidak seimbang. Sebagian penelitian terdahulu hanya menerapkan satu algoritma klasifikasi tanpa melakukan evaluasi komparatif, sementara penelitian lainnya belum mengintegrasikan teknik penanganan *class imbalance* yang dapat memengaruhi kemampuan model dalam mengenali pelanggan yang berpotensi merespon penawaran asuransi. Oleh karena itu, penelitian ini bertujuan untuk membandingkan kinerja Logistic Regression dan Random Forest dalam memprediksi respon pelanggan asuransi kendaraan dengan menerapkan *Synthetic Minority Over-sampling Technique* (SMOTE) pada data latih. Kebaruan penelitian ini terletak pada evaluasi komparatif kedua algoritma klasifikasi tersebut pada kasus prediksi respon pelanggan asuransi kendaraan dengan data yang tidak seimbang menggunakan teknik SMOTE. Adapun kontribusi penelitian ini adalah memberikan bukti empiris mengenai efektivitas masing-masing algoritma dalam menangani data yang tidak seimbang melalui evaluasi menggunakan berbagai metrik klasifikasi, sehingga dapat menjadi referensi dalam pemilihan model prediksi yang lebih tepat untuk mendukung strategi pemasaran perusahaan asuransi.

2. METODE PENELITIAN

Penelitian ini dilakukan untuk memprediksi respon pelanggan terhadap produk asuransi kendaraan menggunakan pendekatan klasifikasi dengan mempertimbangkan permasalahan ketidakseimbangan data. Penelitian ini menggunakan dataset *Vehicle Insurance* yang diperoleh dari platform Kaggle yang berisi informasi terkait karakteristik pelanggan seperti usia, jenis kelamin, status kepemilikan asuransi sebelumnya (*Previously_Insured*), kondisi kendaraan (*Vehicle_Damage*), serta atribut lainnya yang relevan terhadap respon pelanggan. Dataset ini banyak digunakan dalam studi analisis perilaku pelanggan pada industri asuransi karena merepresentasikan kondisi nyata dalam pengambilan keputusan pemasaran.

Penelitian diawali dengan tahap *preprocessing* data untuk memperbaiki kualitas data sebelum digunakan dalam proses pemodelan. Tahap ini meliputi penghapusan atribut yang tidak relevan, transformasi data kategorikal menjadi numerik menggunakan teknik *encoding*, serta pemisahan antara variabel fitur dan variabel target. Variabel target dalam penelitian ini adalah respon pelanggan terhadap penawaran asuransi, yang diklasifikasikan ke dalam dua kelas, yaitu merespon dan tidak merespon.

Setelah tahap preprocessing selesai dilakukan, dataset kemudian dibagi menjadi data pelatihan (*training set*) dan data pengujian (*testing set*) menggunakan metode *train-test split* dengan proporsi 80% data pelatihan dan 20% data pengujian. Proses pembagian data dilakukan menggunakan parameter random_state = 42 untuk menjaga konsistensi hasil eksperimen serta menerapkan teknik *stratified sampling* melalui parameter *stratify* guna mempertahankan proporsi distribusi kelas pada data pelatihan dan data pengujian. Data pelatihan digunakan untuk proses penanganan ketidakseimbangan data, pelatihan model, dan *hyperparameter tuning*, sedangkan data pengujian digunakan sebagai data independen untuk mengevaluasi performa akhir model. Secara matematis, proses pembagian data dapat dinyatakan melalui persamaan (1)

$$Data = d_{train} + d_{test} \quad (1)$$

dengan:

d_{train} : 80% data pelatihan
d_{test} : 20% data pengujian

Permasalahan utama pada dataset ini adalah ketidakseimbangan distribusi kelas (*class imbalance*), di mana jumlah pelanggan yang tidak merespons penawaran asuransi jauh lebih banyak dibandingkan pelanggan yang merespons. Untuk mengatasi kondisi tersebut, penelitian ini menerapkan metode SMOTE pada data pelatihan. Metode ini menghasilkan sampel sintetis pada kelas minoritas berdasarkan pendekatan tetangga terdekat (*k-nearest neighbors*), sehingga distribusi data menjadi lebih seimbang dan model dapat mengenali pola pada kelas minoritas dengan lebih baik[14]. Secara umum, proses pembentukan sampel sintetis pada SMOTE dapat dinyatakan dengan persamaan (2)

$$x_{new} = x_i + \text{rand}(0,1) + (x_{nn} - x_i) \quad (2)$$

dengan:

x_i : data minoritas asli
x_nn : tetangga terdekat dari data minoritas
rand(0,1) : bilangan acak antara 0 dan 1

Pada penelitian ini, teknik SMOTE diterapkan hanya pada data pelatihan untuk menghindari terjadinya *data leakage* yang dapat menyebabkan hasil evaluasi model menjadi bias. Implementasi SMOTE menggunakan parameter random_state = 42 dan nilai *default* k_neighbors = 5. Parameter k_neighbors menentukan jumlah tetangga terdekat yang digunakan dalam proses pembangkitan sampel sintetis pada kelas minoritas. Setelah proses *oversampling* dilakukan, distribusi data pada kelas minoritas menjadi lebih seimbang sehingga model dapat mempelajari karakteristik kedua kelas secara lebih optimal.

Setelah proses penyeimbangan data selesai dilakukan, tahap berikutnya adalah proses normalisasi data menggunakan *StandardScaler* khusus pada model Logistic Regression. Normalisasi dilakukan agar seluruh fitur memiliki skala yang seragam sehingga dapat meningkatkan stabilitas dan performa model. Rumus standardisasi dinyatakan pada persamaan (3)

$$z = \frac{(x - \mu)}{\sigma} \quad (3)$$

dengan:

x : nilai data
μ : rata-rata data
σ : standar deviasi

Proses pemodelan dilakukan dengan menerapkan dua algoritma klasifikasi, yaitu Logistic Regression dan Random Forest. Logistic Regression dipilih sebagai model baseline karena memiliki struktur yang sederhana, efisien, serta mudah dipahami dalam proses interpretasi hasil prediksi. Model ini memprediksi probabilitas kelas menggunakan fungsi sigmoid persamaan (4).

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \quad (4)$$

Di sisi lain, Random Forest dipilih karena mampu mengenali pola hubungan yang lebih kompleks antar fitur dan meningkatkan kualitas klasifikasi melalui konsep ensemble learning. Metode ini bekerja dengan membentuk sejumlah pohon keputusan (decision tree) yang kemudian menghasilkan prediksi akhir berdasarkan hasil voting terbanyak. Hal ini dituliskan melalui persamaan (5)

$$\hat{y} = \text{model}(T_1(x), T_2(x), \ldots T_n(x)) \quad (5)$$

Dengan:

T_1, T_2, T_n : pohon keputusan dalam random forest

Mode : kelas dengan jumlah voting terbanyak

Untuk memperoleh performa model yang optimal, dilakukan proses *hyperparameter tuning* pada algoritma Random Forest menggunakan metode RandomizedSearchCV. Parameter yang dievaluasi meliputi jumlah pohon keputusan (n_estimators = {50,100}), kedalaman maksimum pohon (max_depth = {5,10}), dan jumlah minimum sampel untuk melakukan pemisahan node (min_samples_split = {2,5}). Dari total delapan kemungkinan kombinasi parameter, RandomizedSearchCV melakukan eksplorasi sebanyak tiga kombinasi secara acak (n_iter = 3) sehingga proses pencarian parameter terbaik dapat dilakukan secara lebih efisien. Proses *tuning* menggunakan *cross-validation* sebanyak 3-fold dengan metrik evaluasi *F1-score*. Parameter terbaik yang diperoleh kemudian digunakan untuk membangun model Random Forest yang dievaluasi menggunakan data pengujian.

Tahap terakhir penelitian dilakukan dengan mengevaluasi performa model menggunakan beberapa metrik klasifikasi seperti *confusion matrix*, *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Evaluasi tersebut digunakan untuk menilai kemampuan model dalam mengidentifikasi pola klasifikasi pada data dengan distribusi kelas yang tidak seimbang. Rumus evaluasi yang digunakan meliputi persamaan (6), (7), (8).

$$\text{precission} = \frac{TP}{(TP + FP)} \quad (6)$$

$$\text{recall} = \frac{TP}{(TP + FN)} \quad (7)$$

$$\text{F1-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (8)$$

dengan:

TP = True Positive

FP = False Positive

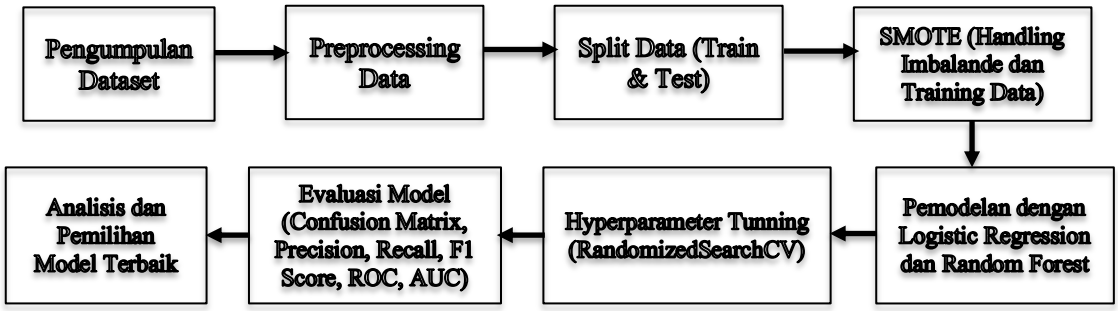
FN = False Negative

Selanjutnya, performa model dianalisis berdasarkan hasil evaluasi untuk menentukan model terbaik yang mampu memberikan keseimbangan antara akurasi dan kemampuan deteksi terhadap kelas minoritas. Model terbaik tersebut diharapkan dapat digunakan sebagai dasar pendukung pengambilan keputusan pada strategi pemasaran produk asuransi kendaraan. Adapun keseluruhan tahapan penelitian ini dirangkum pada Gambar 1.

3. HASIL DAN PEMBAHASAN

Pada tahap awal analisis, dilakukan eksplorasi terhadap distribusi data untuk memahami karakteristik variabel target yang digunakan. Berdasarkan hasil analisis menunjukkan bahwa data respon pelanggan terhadap produk asuransi kendaraan memiliki distribusi yang tidak seimbang, di mana jumlah pelanggan yang tidak merespon jauh lebih besar dibandingkan dengan pelanggan

yang merespon. Ketidakseimbangan distribusi ini terlihat cukup signifikan, di mana kelas tidak merespon (kelas 0) mendominasi sebagian besar dataset dibandingkan dengan kelas merespon (kelas 1). Kondisi ini mengindikasikan bahwa dataset memiliki karakteristik *imbalanced*, yang berpotensi menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas. Akibatnya, model dapat menghasilkan akurasi yang tinggi secara keseluruhan, namun memiliki kemampuan yang rendah dalam mendeteksi pelanggan yang benar-benar memiliki potensi untuk merespon. Adapun distribusi jumlah data pada masing-masing kelas ditunjukkan pada Tabel 1.



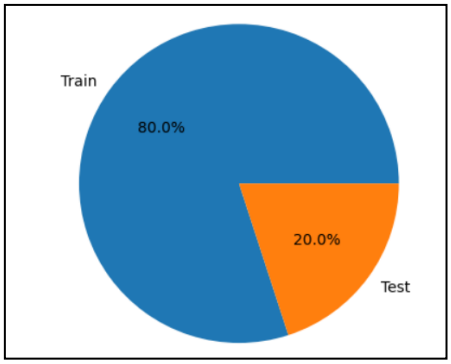
Gambar 1. Tahapan penelitian

Berdasarkan Tabel 1, dapat dilihat bahwa proporsi data antara kelas mayoritas dan minoritas cukup tidak seimbang, sehingga diperlukan penanganan khusus pada tahap selanjutnya untuk meningkatkan keseimbangan data dan performa model klasifikasi.

Tabel 1. Distribusi data sebelum SMOTE

Kelas Respon	Jumlah Data	Persentase
Tidak Respon (0)	255642	83.61%
Respon (1)	50081	16.39%
Total	305723	100%

Pada tahap *preprocessing*, seluruh atribut kategorikal seperti *Gender*, *Vehicle_Age*, dan *Vehicle_Damage* dikonversi ke dalam bentuk numerik melalui proses *encoding* agar dapat diproses oleh model. Selain itu, tidak ditemukan adanya *missing values* sehingga seluruh data dapat digunakan dalam proses pelatihan model. Setelah proses transformasi selesai, dataset dibagi menjadi data *training* dan data *testing* menggunakan metode *train-test split*. Pembagian ini bertujuan untuk mengevaluasi kemampuan model dalam melakukan prediksi terhadap data yang belum pernah digunakan sebelumnya. Proporsi pembagian data terdiri dari 80% data *training* dan 20% data *testing*. Distribusi pembagian dataset tersebut ditunjukkan pada Gambar 2.



Gambar 2. Distribusi data training dan testing (80:20)

Berdasarkan gambar 2, terlihat bahwa sebagian besar data digunakan untuk proses pelatihan model, sedangkan sebagian lainnya digunakan untuk proses evaluasi. Pembagian ini

diharapkan dapat menghasilkan model yang memiliki kemampuan prediksi yang baik serta tidak mengalami overfitting.

Pada tahapan selanjutnya akan dilakukan penanganan terhadap ketidakseimbangan data (*class imbalance*) yang didominasi kelas yang tidak merespon (0) bila dibandingkan kelas yang merespon (1), sehingga berpotensi menyebabkan model klasifikasi menjadi bias terhadap kelas yang mayoritas. Untuk mengatasi permasalahan tersebut, akan diterapkan teknik SMOTE pada data pelatihan. SMOTE akan bekerja dengan menghasilkan data sintetis pada kelas minoritas sehingga jumlah data pada kedua kelas menjadi lebih seimbang[15]. Penerapan teknik ini bertujuan untuk meningkatkan kemampuan model dalam mengenali pola pada kelas minoritas, yang dalam konteks penelitian ini merupakan pelanggan yang berpotensi merespon penawaran asuransi.

Pada penelitian ini, teknik SMOTE diterapkan hanya pada data pelatihan (*training* data) untuk menghindari terjadinya *data leakage* yang dapat menyebabkan hasil evaluasi model menjadi bias. Implementasi SMOTE menggunakan parameter random_state = 42 dan k_neighbors = 5, di mana nilai k_neighbors merupakan jumlah tetangga terdekat yang digunakan dalam proses pembangkitan sampel sintetis pada kelas minoritas. Pemilihan nilai tersebut mengikuti konfigurasi default yang umum digunakan pada implementasi SMOTE. Setelah proses *oversampling* dilakukan, jumlah data pada kelas minoritas meningkat hingga seimbang dengan kelas mayoritas sebelum dilakukan proses pelatihan model. Hasil penerapan SMOTE menunjukkan perubahan terhadap distribusi data. Sebelum penerapan SMOTE, jumlah data pada kelas tidak merespon (kelas 0) sebanyak 255642 data, sedangkan kelas merespon (kelas 1) hanya sebanyak 50081 data. Setelah penerapan SMOTE, jumlah data pada kedua kelas menjadi seimbang, yaitu masing-masing sebanyak 255642 data. Perbandingan distribusi data sebelum dan sesudah penerapan SMOTE ditunjukkan pada Tabel 2.

Tabel 2. Perbandingan distribusi data sebelum dan sesudah penerapan SMOTE

Kelas Respon	Sebelum SMOTE	Setelah SMOTE
Tidak Respon (0)	255642	255642
Respon (1)	50081	255642

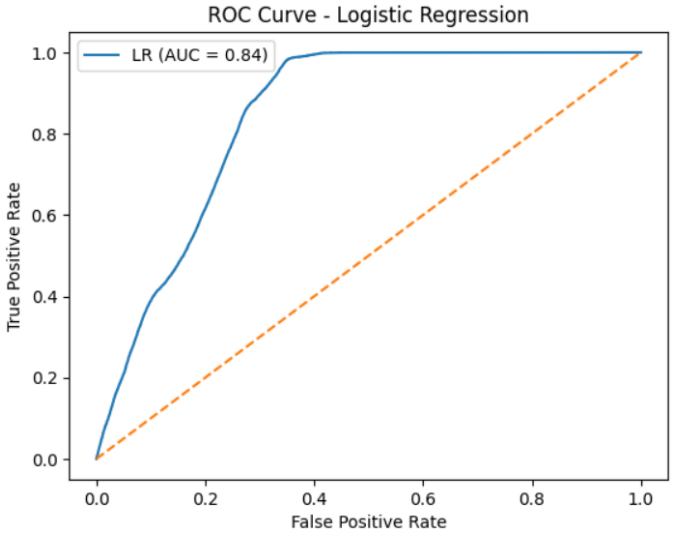
Setelah dilakukan proses penyeimbangan data menggunakan teknik SMOTE, maka tahap selanjutnya adalah pemodelan menggunakan algoritma Logistic Regression. Model ini digunakan sebagai *baseline* untuk mengukur performa awal dalam memprediksi respon pelanggan terhadap produk asuransi kendaraan.

Tabel 3. Hasil Confusion Matrix pada Model Logistic Regression

	Precision	Recall	F1-score	Support
Negatif (0)	0,99	0,65	0,79	63911
Positif (1)	0,35	0,98	0,52	12520
Accuracy			0,70	76431

Berdasarkan hasil pengujian Logistic Regression pada Tabel 3, model menunjukkan kemampuan yang sangat tinggi dalam mengenali pelanggan yang memberikan respon, yang ditunjukkan oleh nilai *recall* kelas positif sebesar 0,98. Hasil tersebut mengindikasikan bahwa sebagian besar pelanggan yang benar-benar merespon berhasil terdeteksi oleh model. Namun, nilai *precision* pada kelas positif masih tergolong rendah, yaitu 0,35. Kondisi ini menunjukkan bahwa masih terdapat cukup banyak prediksi positif yang sebenarnya berasal dari pelanggan yang tidak merespon.

Secara umum, Logistic Regression menghasilkan nilai akurasi sebesar 0,70 dengan *F1-score* kelas positif sebesar 0,52. Hasil tersebut menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam mendeteksi kelas minoritas, meskipun keseimbangan antara nilai *precision* dan *recall* masih belum optimal. Selain itu, kurva ROC pada Gambar 3 memperlihatkan bahwa model memiliki kemampuan yang cukup baik dalam membedakan kelas positif dan negatif dengan nilai AUC sebesar 0,84.



Gambar 3. ROC Curve untuk Model Logistic Regression

Selanjutnya akan dilakukan pemodelan menggunakan Random Forest, di mana model ini digunakan untuk mengatasi keterbatasan model linier dalam menangkap hubungan kompleks antar fitur, khususnya pada data dengan pola non-linear.

Berdasarkan hasil proses *hyperparameter tuning* menggunakan metode RandomizedSearchCV, diperoleh parameter terbaik yang menghasilkan performa optimal, di antaranya jumlah pohon (*n_estimators*) sebesar 100, kedalaman maksimum pohon (*max_depth*) sebesar 10, serta parameter *min_samples_split* sebesar 2. Kombinasi parameter tersebut memberikan keseimbangan yang baik antara kemampuan model dalam mempelajari pola data dan kemampuan generalisasi terhadap data baru.

Hasil evaluasi Random Forest menunjukkan performa yang lebih baik dibandingkan Logistic Regression. Model ini memperoleh akurasi sebesar 0,80 dengan *precision* kelas positif sebesar 0,44 dan *recall* sebesar 0,87. Selain itu, *F1-score* sebesar 0,59 menunjukkan bahwa Random Forest mampu menghasilkan keseimbangan *precision* dan *recall* yang lebih baik dibandingkan Logistic Regression. Adapun hasil *confusion matrix* model Random Forest ditampilkan pada Tabel 4.

Tabel 4. Hasil Confusion Matrix pada Model Random Forest

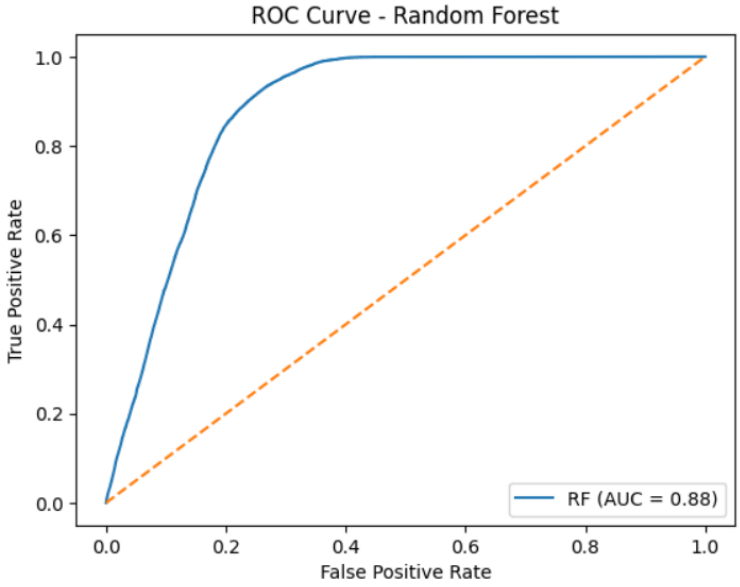
	Precision	**Recall**	**F1-score**	**Support**
Negatif (0)	0,97	0,78	0,87	63911
Positif (1)	0,44	0,87	0,59	12520
Accuracy			0,80	76431

Selain itu, nilai ROC-AUC sebesar 0,88 pada Gambar 4 menunjukkan bahwa Random Forest memiliki kemampuan yang lebih baik dalam membedakan antara kelas respon dan tidak respon.

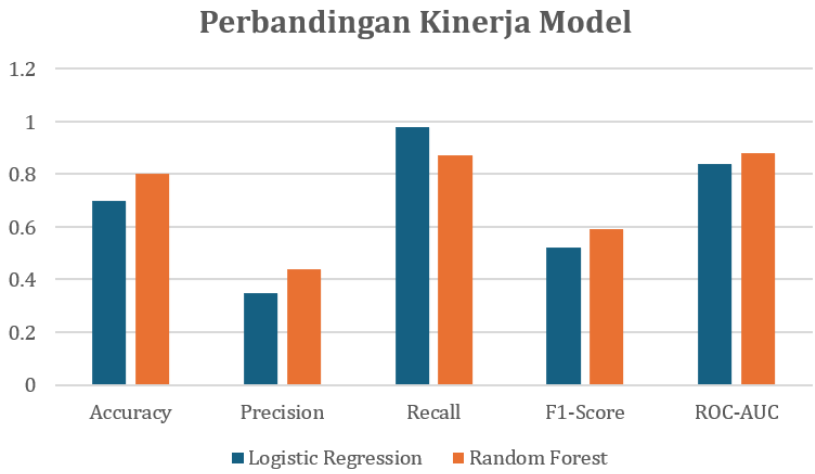
Selanjutnya untuk mengetahui model terbaik, dilakukan perbandingan kinerja antara Logistic Regression dan Random Forest berdasarkan beberapa metrik evaluasi. Hasil perbandingan ditunjukkan pada Gambar 5.

Berdasarkan Gambar 5, Random Forest menunjukkan performa yang lebih unggul dibandingkan Logistic Regression pada sebagian besar metrik evaluasi, khususnya *accuracy*, *precision*, *F1-score*, dan *ROC-AUC*. Hal ini menunjukkan bahwa Random Forest mampu menghasilkan prediksi yang lebih akurat dan seimbang dalam membedakan antara pelanggan yang merespon dan tidak merespon. Sebaliknya, Logistic Regression memiliki nilai *recall* yang sangat tinggi, namun diikuti dengan *precision* yang rendah, yang mengindikasikan bahwa model cenderung menghasilkan banyak *false positive*. Kondisi ini menunjukkan bahwa Logistic

Regression lebih agresif dalam mendeteksi pelanggan potensial, tetapi kurang efisien dalam konteks penentuan target pemasaran.



Gambar 4. ROC Curve untuk Model Random Forest



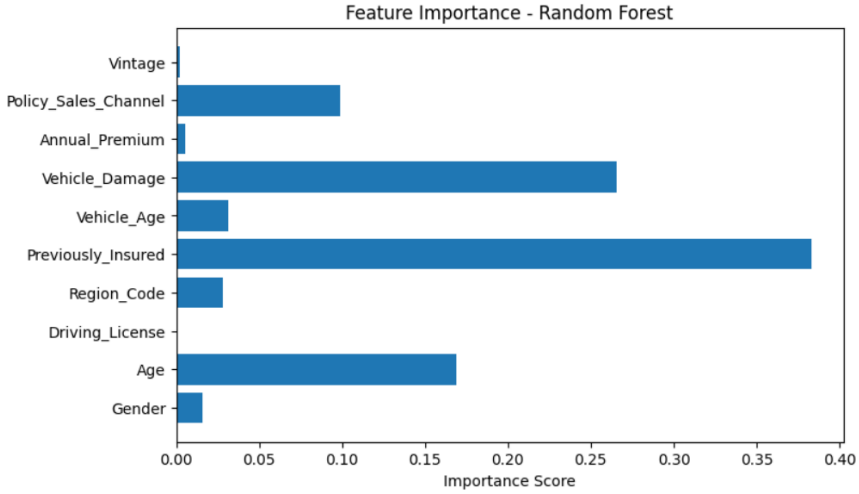
Gambar 5. Grafik perbandingan kinerja model

Perbedaan performa kedua model tersebut disebabkan karena Logistic Regression merupakan model klasifikasi linear yang bekerja dengan membangun hubungan linear antara variabel independen dan probabilitas kelas target [7]. Pendekatan ini relatif sederhana dan mudah diinterpretasikan, namun memiliki keterbatasan dalam menangkap hubungan yang kompleks dan non-linear antar fitur. Sebaliknya, Random Forest merupakan metode *ensemble* yang menggabungkan banyak pohon keputusan sehingga mampu memodelkan pola data yang lebih kompleks serta menangkap interaksi antar variabel secara lebih efektif [8]. Karakteristik tersebut memungkinkan Random Forest menghasilkan nilai *precision*, *F1-score*, dan ROC-AUC yang lebih tinggi dibandingkan Logistic Regression pada dataset pelanggan asuransi kendaraan yang memiliki karakteristik data yang beragam.

Secara keseluruhan, Random Forest lebih direkomendasikan karena mampu menjaga keseimbangan antara *precision* dan *recall*, yang tercermin pada nilai *F1-score* dan ROC-AUC yang lebih tinggi. Hal ini menunjukkan bahwa Random Forest tidak hanya mampu mendeteksi

pelanggan potensial, tetapi juga lebih selektif dalam mengurangi kesalahan prediksi, sehingga lebih sesuai untuk diimplementasikan dalam strategi pemasaran berbasis data.

Selanjutnya untuk memahami faktor-faktor yang paling berpengaruh dalam prediksi respon pelanggan, dilakukan analisis *feature importance* menggunakan model Random Forest. Hasil analisis menunjukkan bahwa beberapa fitur memiliki kontribusi yang dominan dalam menentukan prediksi model. Fitur yang paling berpengaruh adalah *Previously_Insured*, diikuti oleh *Vehicle_Damage* dan *Age*. Tingginya kontribusi fitur *Previously_Insured* menunjukkan bahwa status kepemilikan asuransi sebelumnya merupakan faktor utama dalam menentukan kemungkinan pelanggan untuk merespon penawaran asuransi.



Gambar 6. Feature importance RF

Berdasarkan Gambar 6 pelanggan yang belum memiliki asuransi sebelumnya cenderung memiliki peluang lebih tinggi untuk merespon penawaran. Selain itu, pelanggan yang pernah mengalami kerusakan kendaraan (*Vehicle_Damage*) juga menunjukkan kecenderungan yang lebih tinggi dalam mempertimbangkan asuransi. Faktor usia (*Age*) juga turut memberikan pengaruh yang cukup signifikan dalam menentukan perilaku pelanggan.

Pembahasan

Hasil penelitian menunjukkan bahwa Random Forest menghasilkan performa yang lebih baik dibandingkan Logistic Regression dengan nilai akurasi sebesar 0,80, F1-score sebesar 0,59, dan ROC-AUC sebesar 0,88. Temuan ini sejalan dengan penelitian [6] yang menunjukkan bahwa Random Forest memiliki kemampuan klasifikasi yang lebih baik dibandingkan Logistic Regression pada kasus prediksi berbasis data pelanggan. Keunggulan tersebut disebabkan oleh kemampuan Random Forest dalam membangun banyak pohon keputusan dan menggabungkan hasil prediksi melalui mekanisme *ensemble learning* sehingga mampu mengurangi variansi model dan meningkatkan kemampuan generalisasi. Perbedaan performa antara kedua model juga dapat dijelaskan dari karakteristik algoritmanya. Logistic Regression merupakan model linear yang mengasumsikan hubungan linear antara variabel independen dan probabilitas kelas target [7]. Pada data pelanggan asuransi, hubungan antar variabel seperti usia pelanggan, status kepemilikan asuransi sebelumnya, kondisi kendaraan, dan kerusakan kendaraan memungkinkan memiliki pola yang kompleks dan non-linear. Kondisi ini menyebabkan Logistic Regression mengalami keterbatasan dalam memodelkan hubungan antar fitur sehingga menghasilkan *precision* yang relatif rendah. Sebaliknya, Random Forest mampu menangkap interaksi dan pola non-linear secara lebih efektif melalui struktur pohon keputusan yang dimilikinya [8].

Hasil penelitian juga menunjukkan bahwa Logistic Regression memperoleh nilai *recall* yang sangat tinggi sebesar 0,98 namun memiliki *precision* yang rendah sebesar 0,35. Kondisi ini menunjukkan bahwa model mampu mendeteksi hampir seluruh pelanggan yang benar-benar memberikan respon, tetapi menghasilkan jumlah *false positive* yang cukup tinggi. Dalam konteks

pemasaran asuransi, kondisi tersebut dapat menyebabkan perusahaan melakukan promosi kepada banyak pelanggan yang sebenarnya tidak memiliki minat terhadap produk yang ditawarkan. Sebaliknya, Random Forest menghasilkan *precision* yang lebih tinggi sebesar 0,44 dengan *recall* sebesar 0,87 sehingga memberikan keseimbangan yang lebih baik antara kemampuan mendeteksi pelanggan potensial dan meminimalkan kesalahan prediksi.

Selain pemilihan algoritma, penerapan SMOTE juga memberikan kontribusi penting terhadap peningkatan performa model. Menurut penelitian [3] dan [13] bahwa ketidakseimbangan distribusi kelas merupakan salah satu faktor utama yang menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas. Pada penelitian ini, sebelum penerapan SMOTE, proporsi pelanggan yang merespon hanya sebesar 16,39% dari keseluruhan data. Setelah proses *oversampling* dilakukan, distribusi kedua kelas menjadi seimbang sehingga model dapat mempelajari karakteristik kelas minoritas dengan lebih baik. Hasil tersebut ternyata mendukung hasil penelitian [15] yang menyatakan bahwa teknik berbasis SMOTE mampu meningkatkan kemampuan model dalam mengenali kelas minoritas pada permasalahan data tidak seimbang. Temuan penelitian ini juga memperkuat hasil penelitian [12] yang menunjukkan bahwa penanganan class *imbalance* merupakan faktor penting dalam pengembangan model prediksi pada industri asuransi. Dengan distribusi data yang lebih seimbang, model mampu memberikan performa yang lebih stabil dan tidak hanya berfokus pada kelas mayoritas. Oleh karena itu, kombinasi Random Forest dan SMOTE dapat menjadi pendekatan yang efektif untuk mendukung pengambilan keputusan pemasaran yang lebih tepat sasaran pada perusahaan asuransi kendaraan.

4. KESIMPULAN

Berdasarkan hasil penelitian, penerapan teknik SMOTE terbukti mampu mengatasi ketidakseimbangan data sehingga meningkatkan kemampuan model dalam mendeteksi kelas minoritas. Hasil perbandingan menunjukkan bahwa Random Forest memiliki kinerja yang lebih baik dibandingkan Logistic Regression, yang ditunjukkan oleh nilai akurasi, *F1-score*, dan ROC-AUC yang lebih tinggi serta keseimbangan yang lebih baik antara *precision* dan *recall*. Sementara itu, Logistic Regression memiliki keunggulan pada nilai *recall* yang tinggi, namun cenderung menghasilkan lebih banyak *false positive*. Analisis *feature importance* menunjukkan bahwa variabel *Previously_Insured*, *Vehicle_Damage*, dan *Age* merupakan faktor yang paling berpengaruh terhadap respon pelanggan. Dengan demikian, kombinasi Random Forest dan SMOTE dapat direkomendasikan sebagai pendekatan yang efektif untuk mendukung strategi pemasaran pada industri asuransi kendaraan.

Penelitian ini memiliki keterbatasan yaitu hanya membandingkan dua algoritma klasifikasi dan menggunakan satu teknik penanganan data tidak seimbang pada satu dataset. Penelitian selanjutnya dapat mengembangkan studi ini dengan membandingkan algoritma lain seperti XGBoost, LightGBM, atau CatBoost serta mengevaluasi teknik penanganan *class imbalance* yang berbeda untuk meningkatkan performa prediksi.

5. SARAN

Penelitian selanjutnya disarankan untuk mengeksplorasi algoritma klasifikasi lain yang lebih kompleks atau berbasis *ensemble* guna meningkatkan performa model. Selain itu, penggunaan metode penanganan data tidak seimbang selain SMOTE, seperti ADASYN atau pendekatan *cost-sensitive learning*, perlu dikaji lebih lanjut untuk mengetahui efektivitasnya. Pengembangan juga dapat dilakukan dengan menambahkan fitur yang lebih representatif agar model memiliki kemampuan generalisasi yang lebih baik

DAFTAR PUSTAKA

[1] U. S. Sulistyawati and M. Munawir, “Decoding Big Data : Mengubah Data Menjadi Keunggulan Kompetitif dalam Pengambilan Keputusan Bisnis Abstrak,” *J. Manaj. dan*

Teknol., vol. 1, no. 2, p. 14, 2024, doi: 10.63447/jmt.v1i2.1114.
[2] N. K. Wahyu Utami and M. I. Padli Nasution, "Dampak Penerapan Big Data dalam Sistem Informasi Manajemen Terhadap Prediksi Tren dan Strategi Pemasaran Konsumen di Indonesia Tahun 2025," *Int. J. Islam. Bus. Manag.*, vol. 4, no. 6, p. 8, 2025.
[3] M. Altalhan, A. Algarni, and M. T. Alouane, "Imbalanced Data Problem in Machine Learning : A Review," *IEEE Access*, vol. 13, no. December 2024, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.
[4] B. Van Giffen, D. Herhausen, and T. Fahse, "Overcoming the pitfalls and perils of algorithms : A classification of machine learning biases and mitigation methods," *J. Bus. Res.*, vol. 144, no. January, pp. 93–106, 2022, doi: 10.1016/j.jbusres.2022.01.076.
[5] R. Graf, M. Zeldovich, and S. Friedrich, "Comparing linear discriminant analysis and supervised learning algorithms for binary classification — A method comparison study," *Biometrical J.*, no. March, p. 20, 2022, doi: 10.1002/bimj.202200098.
[6] T. Wahyuningsih, D. Manongga, I. Sembiring, and S. Wijono, "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Argument," *Procedia Comput. Sci.*, vol. 234, pp. 349–356, 2024, doi: 10.1016/j.procs.2024.03.014.
[7] D. Dey *et al.*, "The proper application of logistic regression model in complex survey data : a systematic review," *BMC Med. Res. Methodol.*, vol. 5, p. 18, 2025, doi: 10.1186/s12874-024-02454-5.
[8] P. Das and D. A. S. Kironmala, "Machine Learning - Based Rainfall Forecasting with Multiple Non - Linear Feature Selection Algorithms," *Water Resour. Manag.*, no. October, p. 29, 2022, doi: 10.1007/s11269-022-03341-8.
[9] E. Hendrawan, D. Zakaria, E. Salwa, and J. Heikal, "Customer Renewal Prediction for Motor Vehicle Insurance Using Binary Logistic Regression in PT XYZ Insurance," *Innov. J. Soc. Sci. Res.*, vol. 4, no. 6, p. 10, 2024, doi: 10.31004/innovative.v4i6.16478.
[10] A. S. Honggowibowo, M. K. Nasrillah, D. Nugraheny, and N. D. Retnowati, "Sistem Rekomendasi Asuransi Mobil Berbasis Web dengan Pendekatan Weighted Product," *Indones. J. Comput. Sci.*, vol. 4, no. 1, p. 7, 2025, doi: 10.31294/3r2c2857.
[11] Q. A. Siregar, R. Meliyani, H. Rahmah, M. H. Musito, M. Amelia, and C. Secu, "Prediksi Risiko Gagal Bayar Premi Menggunakan Algoritma Gradient Boosting: Studi Travel Insurance Prediction," *JUKOMTEK (Jurnal Komput. dan Teknol.*, vol. 4, no. 1, p. 4, 2024, doi: 10.64626/jukomtek.v3i2.565.
[12] F. Khamesian, M. Esna-ashari, E. D. Ofosu-hene, and F. Khanizadeh, "Risk Classification of Imbalanced Data for Car Insurance Companies : Machine Learning Approaches," *Int. J. Math. Model. Comput.*, vol. 12, no. 03, p. 10, 2022, doi: 10.30495/ijm2c.2022.1958403.1252.
[13] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, 2024, doi: 10.1007/s10994-022-06268-8.
[14] S. Mohammad, H. Mirsadeghi, H. Bahsi, R. Vaarandi, and W. Inoubli, "Learning From Few Cyber-Attacks : Addressing the Class Imbalance Problem in Machine Learning-Based Intrusion Detection in Software-Defined Networking," *IEEE Access*, vol. 11, no. December, 2023, doi: 10.1109/ACCESS.2023.3341755.
[15] D. Dablain, B. Krawczyk, and N. V Chawla, "DeepSMOTE : Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 9, p. 15, 2021, doi: 10.1109/TNNLS.2021.3136503.