JURNAL FASILKOM
Volume 16 No. 1 | April 2026: 120-126

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Prediksi *Lead Scoring* untuk Optimasi Penjualan Menggunakan *Random Forest* dan Teknik *SMOTE*

Daffa Pratama Putra[1], M. Rizki Al Akbar[2], Dimas Agil Kusuma[3], Ali Ibrahim[4], Fathoni[5]
[2]Program Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya
[1]09031282328090@student.unsri.ac.id*, [2]09031282328087@student.unsri.ac.id, [3]09031282328054@student.unsri.ac.id,
[4]aliibrahim210784@gmail.com, [5]fathoni@unsri.ac.id

Abstract

Accurate lead scoring systems have become a strategic necessity for organizations operating in data-driven marketing environments, as they enable systematic identification of high-value customer prospects to maximize sales conversion efficiency. A fundamental challenge confronting conventional classification models is the class imbalance inherent in real-world marketing data, which induces majority-class bias and substantially reduces sensitivity toward minority-class prospects. This study proposes a Random Forest (RF)-based lead scoring prediction model integrated with the Synthetic Minority Over-sampling Technique (SMOTE) to address this limitation systematically. The dataset employed is the Lead Scoring Dataset from Kaggle, comprising 9,240 customer prospect records from an educational company with a class imbalance ratio of 1.59:1. Preprocessing included missing value treatment, removal of attributes exceeding 40% data loss, mode-based imputation, and categorical feature encoding. Following an 80:20 stratified split, SMOTE was applied exclusively to the training set to produce a balanced class distribution and prevent data leakage. The RF model was configured with n_estimators = 100, max_features = 'sqrt', and class_weight = 'balanced'. The proposed RF+SMOTE model achieved accuracy of 88.80%, precision of 86.44%, recall of 84.13%, F1-Score of 85.27%, and AUC-ROC of 0.9453, outperforming the baseline across four of five evaluation metrics. The most notable improvement was observed in recall, with a gain of 1.26 percentage points. Stratified 5-Fold Cross-Validation confirmed robust generalization capability, with AUC-ROC values consistently ranging between 94% and 95%. These findings demonstrate that the hybrid RF+SMOTE approach effectively enhances high-potential prospect detection while maintaining overall model stability for real-world Customer Relationship Management (CRM) deployment.

Keywords: Lead Scoring, Random Forest, SMOTE, Class Imbalance, Customer Relationship Management

Abstrak

Sistem *lead scoring* yang akurat merupakan kebutuhan strategis dalam pemasaran berbasis data, mengingat kemampuannya mengidentifikasi prospek pelanggan bernilai tinggi guna memaksimalkan efisiensi konversi penjualan. Permasalahan utama yang dihadapi model klasifikasi konvensional adalah ketidakseimbangan distribusi kelas (*class imbalance*) pada data pemasaran nyata, yang menyebabkan model cenderung bias terhadap kelas mayoritas dan gagal mendeteksi prospek potensial dari kelas minoritas. Penelitian ini mengusulkan model prediksi *lead scoring* berbasis *Random Forest* (RF) yang diintegrasikan dengan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi permasalahan tersebut secara sistematis. Dataset yang digunakan adalah *Lead Scoring Dataset* dari Kaggle, terdiri dari 9.240 data prospek pelanggan perusahaan edukasi dengan rasio ketidakseimbangan kelas sebesar 1,59:1. Pra-pemrosesan mencakup penanganan *missing value*, penghapusan atribut dengan kehilangan data lebih dari 40%, imputasi modus, serta *encoding* fitur kategorikal. Setelah pembagian data dengan rasio 80:20, SMOTE diterapkan secara eksklusif pada data latih untuk menghasilkan distribusi kelas yang seimbang dan mencegah *data leakage*. Model RF dikonfigurasi dengan parameter *n_estimators* = 100, *max_features* = 'sqrt', dan *class_weight* = 'balanced'. Model RF+SMOTE yang diusulkan mencapai *accuracy* 88,80%, *precision* 86,44%, *recall* 84,13%, *F1-Score* 85,27%, dan *AUC-ROC* 0,9453, melampaui model *baseline* pada empat dari lima metrik evaluasi. Peningkatan paling signifikan terjadi pada *recall* sebesar 1,26%. Validasi menggunakan *Stratified 5-Fold Cross-Validation* menghasilkan nilai *AUC-ROC* yang konsisten pada rentang 94–95%, mengkonfirmasi kemampuan generalisasi model yang tinggi. Temuan ini membuktikan bahwa pendekatan hibrida RF+SMOTE efektif dalam meningkatkan deteksi prospek potensial sekaligus menjaga stabilitas model untuk implementasi pada sistem *Customer Relationship Management* (CRM) nyata.

Kata kunci: *Lead Scoring, Random Forest, SMOTE, Class Imbalance, Customer Relationship Management*

1. Pendahuluan

Dalam era transformasi digital yang masif, optimalisasi sistem *Customer Relationship Management* (CRM) melalui pemanfaatan big data dan kecerdasan buatan telah menjadi strategi fundamental bagi organisasi global untuk mempertahankan daya saing di pasar yang sangat kompetitif [1]. Fenomena global menunjukkan bahwa perusahaan tidak lagi hanya fokus pada pengumpulan data pelanggan dalam jumlah besar, melainkan pada kemampuan untuk mengekstraksi wawasan prediktif guna meningkatkan efisiensi operasional [2]. Di tingkat internasional, tren riset terkini mulai mengarah pada integrasi model

P-ISSN : 2089-3353
E-ISSN : 2808-9162

pembelajaran mesin yang mampu melakukan *lead prioritization* secara otomatis untuk memastikan tim penjualan fokus pada prospek yang memiliki probabilitas konversi tertinggi [3]. Signifikansi dari otomatisasi *lead scoring* ini terletak pada kemampuannya untuk menjembatani kesenjangan antara strategi pemasaran dan eksekusi penjualan di berbagai sektor industri global [4].

Namun, kendala utama yang sering dihadapi dalam implementasi sistem ini adalah distribusi data pemasaran yang sangat tidak seimbang (*class imbalance*), di mana jumlah pelanggan yang melakukan konversi nyata jauh lebih sedikit dibandingkan dengan prospek yang tidak melakukan pembelian [5]. Ketimpangan kelas ini menjadi masalah serius karena model klasifikasi tradisional cenderung mengalami bias terhadap kelas mayoritas, sehingga sering kali gagal mengidentifikasi "leads" berkualitas yang justru berada pada kelas minoritas [6]. Jika masalah ketidakseimbangan data ini tidak segera diselesaikan, konsekuensinya adalah penurunan akurasi prediksi yang menyebabkan kesalahan alokasi sumber daya penjualan, pemborosan biaya operasional pada prospek yang tidak potensial, serta hilangnya peluang pendapatan yang signifikan bagi perusahaan [7]. Kegagalan model dalam menangani data yang tidak seimbang ini secara langsung akan merusak kepercayaan pemangku kepentingan terhadap sistem pendukung keputusan berbasis data [8].

Beberapa penelitian terdahulu telah menunjukkan efektivitas berbagai algoritma pembelajaran mesin dalam meningkatkan performa klasifikasi pada data CRM yang kompleks. Penelitian terdahulu mengindikasikan bahwa algoritma Random Forest memiliki ketahanan yang luar biasa dalam menangani data tabular pemasaran dan mampu memberikan tingkat akurasi yang stabil dibandingkan model linear konvensional [9]. Selain itu, berbagai studi telah membuktikan bahwa penggunaan teknik *oversampling* seperti *Synthetic Minority Over-sampling Technique* (SMOTE) dapat secara efektif meningkatkan sensitivitas model terhadap kelas minoritas dengan menghasilkan sampel sintetis yang representatif [10]. Penelitian lain juga menegaskan bahwa kombinasi antara teknik penyeimbangan data dan algoritma *ensemble* mampu memberikan hasil yang lebih unggul dalam berbagai domain prediksi risiko dan perilaku pelanggan [11].

Namun, meskipun penggunaan Random Forest dan SMOTE telah banyak dibahas, sebagian besar studi tersebut masih berfokus pada domain keuangan seperti prediksi kebangkrutan atau deteksi penipuan, sementara perhatian terhadap optimasi konversi penjualan dalam konteks *lead scoring* yang dinamis masih sangat terbatas [12]. Meskipun demikian, terdapat keraguan ilmiah mengenai validitas praktis penggunaan SMOTE standar apabila tidak diintegrasikan dengan pemahaman mendalam terhadap karakteristik data pemasaran yang memiliki tingkat derau (noise) tinggi [13]. Eksplorasi terhadap sinkronisasi antara parameter SMOTE dan arsitektur Random Forest guna memitigasi galat prediksi pada data konversi penjualan dengan imbalans ekstrem sejauh ini masih belum banyak dilakukan [14].

Penelitian ini bertujuan untuk merumuskan model prediksi *lead scoring* yang tangguh dengan mengintegrasikan algoritma Random Forest dan teknik SMOTE guna mengatasi hambatan ketidakseimbangan data pada sistem pemasaran. Kontribusi teoretis dari studi ini terletak pada pemetaan parameter optimal untuk penanganan *class imbalance* di domain CRM, sementara secara praktis, model ini memberikan panduan bagi profesional pemasaran dalam meningkatkan rasio konversi penjualan secara signifikan. Nilai kebaruan (*novelty*) yang membedakan penelitian ini dengan riset sebelumnya adalah pendekatan hibrida yang disesuaikan untuk meminimalkan bias pada kelas mayoritas tanpa mengorbankan stabilitas model, sehingga menghasilkan prediksi *lead scoring* yang jauh lebih presisi dan aplikatif untuk kebutuhan bisnis nyata [15].

2. Metode Penelitian

Metode yang digunakan dalam penelitian ini mengikuti tahapan standar *Data Mining*, yang meliputi pengumpulan data, pra-pemrosesan data, pembagian data, penerapan SMOTE, pemodelan, dan evaluasi. Tahapan penelitian ditunjukkan pada Gambar 1 berikut.

Gambar 1. Tahapan Penelitian

2.1. Pengumpulan Data

Penelitian ini menggunakan data sekunder yaitu *Lead Scoring Dataset* yang diperoleh dari platform publik

Author : **Daffa Pratama Putra[1)], M. Rizki Al Akbar[2)], Dimas Agil Kusuma[3)], Ali Ibrahim[4)], Fathoni[5)]**

P-ISSN : 2089-3353
E-ISSN : 2808-9162

Kaggle. Data ini berisi riwayat prospek pelanggan dari sebuah perusahaan edukasi. Dataset ini terdiri dari 9.240 baris data dan 37 kolom atribut. Variabel yang menjadi target prediksi adalah kolom *Converted*, di mana angka '1' berarti prospek berhasil membeli (konversi), dan angka '0' berarti prospek tidak membeli. Berdasarkan pengecekan awal, jumlah prospek yang tidak membeli jauh lebih banyak daripada yang membeli, sehingga terjadi masalah ketidakseimbangan data (*imbalanced data*).

2.2. Pra Pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk membersihkan data mentah agar layak diproses oleh algoritma. Langkah pertama adalah menangani nilai "Select" yang terdapat pada beberapa kolom kategorikal, yang secara sistem berarti pengguna tidak memilih opsi apa pun pada formulir. Nilai ini kemudian diubah menjadi format data kosong (*missing value*). Selanjutnya, kolom atribut yang memiliki persentase data kosong lebih dari 40% dihapus karena dianggap tidak lagi memberikan informasi yang relevan untuk proses prediksi. Untuk sisa data kosong pada kolom yang masih dipertahankan, proses pengisian (*imputasi*) dilakukan menggunakan nilai modus atau nilai yang paling sering muncul. Tahap ini diakhiri dengan proses transformasi data teks menjadi angka (*encoding*) menggunakan *Label Encoding* untuk data dengan dua pilihan dan *One-Hot Encoding* untuk data dengan lebih dari dua pilihan, mengingat algoritma *machine learning* hanya dapat memproses komputasi numerik.

2.3. Pembagian Data dan Penerapan SMOTE

Sebelum data diseimbangkan, dataset terlebih dahulu dibagi menjadi dua bagian, yaitu data latih (training data) sebesar 80% dan data uji (testing data) sebesar 20%. Pemisahan ini dilakukan di awal agar data uji tetap murni dan tidak tercampur. Setelah dibagi, teknik SMOTE (*Synthetic Minority Over-sampling Technique*) diterapkan hanya pada data latih. SMOTE bekerja dengan cara menambahkan data buatan pada kelas minoritas (prospek yang membeli) sehingga jumlahnya seimbang dengan kelas mayoritas (prospek yang tidak membeli). Langkah ini bertujuan untuk memitigasi bias model terhadap kelas mayoritas, sehingga sensitivitas terhadap kelas minoritas dapat ditingkatkan secara optimal.

2.4. Pemodelan Menggunakan *Random Forest*

Himpunan data latih yang distribusinya telah seimbang kemudian digunakan untuk membangun model klasifikasi. Model dikonfigurasi dengan parameter n_*estimators* = 100, *max_features* = 'sqrt', *class_weight* = *'balanced'*, dan random_state = 42 untuk menjamin reprodusibilitas eksperimen. Algoritma yang digunakan dalam penelitian ini adalah *Random Forest*, yang dipilih karena kemampuannya yang stabil dan andal dalam memproses banyak variabel secara bersamaan. Algoritma *ensemble* ini bekerja dengan membangun banyak pohon keputusan (*decision trees*) dan mengambil hasil prediksi terbanyak (*majority voting*) untuk menentukan hasil klasifikasi akhir. Pendekatan ini sangat efektif untuk mencegah terjadinya *overfitting* atau kondisi di mana model terlalu menghafal data latih, sehingga model dapat memprediksi prospek baru dengan akurasi yang lebih baik.

2.5. Evaluasi Model

Mengingat karakteristik data bawaan yang tidak seimbang, penggunaan metrik akurasi secara tunggal dinilai tidak cukup dan dapat memberikan kesimpulan yang keliru. Oleh karena itu, pengujian performa model dilakukan menggunakan Confusion Matrix terhadap himpunan data uji murni untuk mengekstrak nilai True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Evaluasi kinerja pemodelan diukur melalui parameter metrik yang lebih komprehensif, salah satunya adalah Precision untuk mengukur ketepatan tebakan model terhadap prospek yang benar-benar membeli, yang diformulasikan sebagai:

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Selanjutnya, metrik Recall digunakan untuk menilai kemampuan kepekaan model dalam menemukan sebanyak mungkin prospek potensial dari keseluruhan data aktual kelas positif, dengan rumus komputasi:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Guna mendapatkan titik keseimbangan yang optimal antara tingkat presisi dan sensitivitas tebakan, digunakan pula F1-Score yang merupakan nilai rata-rata harmonis dari keduanya. Metrik ini sangat krusial untuk mengevaluasi data pemasaran yang timpang, yang dihitung melalui persamaan:

$$F1 = \frac{2 \times (Precision \times Recall)}{(Precision+Recall)} \quad (3)$$

Sebagai validasi akhir, metrik *Area Under the Receiver Operating Characteristic Curve* (AUC-ROC) juga digunakan untuk memvalidasi kemampuan model secara keseluruhan dalam membedakan probabilitas antara prospek yang sukses dikonversi dan yang gagal.

3. Hasil dan Pembahasan

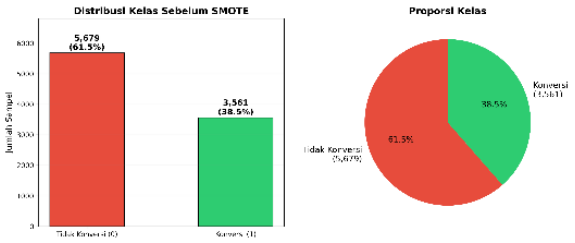
3.1. Distribusi Kelas *Dataset*

Analisis awal terhadap keseluruhan dataset mengungkapkan adanya ketidakseimbangan kelas (*class imbalance*) yang perlu ditangani sebelum proses pemodelan dilakukan. Dari total 9.240 sampel, sebanyak 5.679 sampel (61,5%) termasuk kelas negatif (tidak konversi) dan 3.561 sampel (38,5%) termasuk

Author : **Daffa Pratama Putra[1)], M. Rizki Al Akbar[2)], Dimas Agil Kusuma[3)], Ali Ibrahim[4)], Fathoni[5)]**

P-ISSN : 2089-3353
E-ISSN : 2808-9162

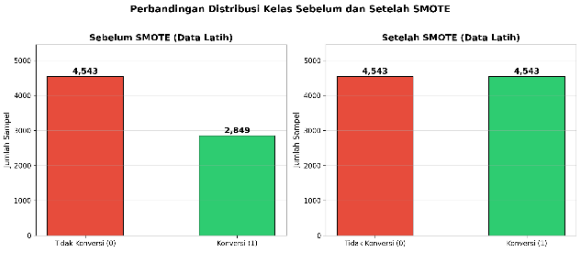
kelas positif (konversi), dengan rasio imbalance sebesar 1,59:1. Distribusi kelas pada keseluruhan dataset divisualisasikan pada Gambar 2. Kondisi ini mengkonfirmasi premis penelitian bahwa data pemasaran secara inheren bersifat tidak seimbang, di mana kelas yang paling bernilai secara bisnis justru merupakan kelas minoritas [5].



Gambar 2. Distribusi Kelas *Dataset Lead Scoring* Sebelum Penerapan SMOTE

3.2. Efektivitas Penerapan SMOTE

Setelah pembagian data dengan rasio 80:20 secara *stratified* (7.392 data latih, 1.848 data uji), teknik SMOTE dengan parameter *k_neighbors* = 5 diterapkan secara eksklusif pada data latih. Penerapan SMOTE secara eksklusif pada data latih merupakan praktik yang krusial untuk mencegah *data leakage* data uji harus tetap merepresentasikan distribusi kelas sesungguhnya agar evaluasi model bersifat objektif [5]. Hasilnya, kelas minoritas pada data latih yang semula berjumlah 2.849 sampel berhasil diseimbangkan menjadi 4.543 sampel, menyamai jumlah kelas mayoritas, sehingga rasio distribusi data latih menjadi 1:1 sempurna dengan total 9.086 sampel. Perubahan distribusi kelas pada data latih sebelum dan setelah penerapan SMOTE ditampilkan secara komparatif pada Gambar 3.



Gambar 3. Distribusi Kelas pada Data Latih Sebelum dan Setelah Penerapan SMOTE

3.3 Evaluasi Performa Model

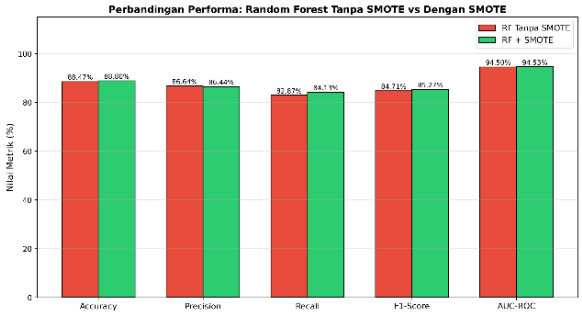
Untuk mengevaluasi kontribusi SMOTE secara kuantitatif, perbandingan performa antara model *Random Forest* tanpa SMOTE sebagai *baseline* dan model yang diusulkan (RF+SMOTE) disajikan secara komprehensif pada Tabel 1. Kedua model dievaluasi pada data uji murni yang terdiri dari 1.848 sampel untuk memastikan objektivitas penilaian.

Tabel 1. Perbandingan Performa Model *Random Forest* Tanpa SMOTE vs RF+SMOTE

Metrik	RF Tanpa SMOTE	RF + SMOTE	Δ
Accuracy	88,47%	88,80%	+0,33%
Precision	86,64%	86,44%	−0,20%
Recall	82,87%	84,13%	+1,26%
F1-Score	84,71%	85,27%	+0,56%
AUC-ROC	94,50%	94,53%	+0,03%

Model RF+SMOTE mencapai *accuracy* 88,80%, *F1-Score* 85,27%, dan *AUC-ROC* 94,53%. Peningkatan paling signifikan secara praktis terjadi pada metrik *Recall*, yakni sebesar 1,26 poin persentase dari 82,87% menjadi 84,13%. Dalam konteks *lead scoring*, *recall* merupakan metrik yang paling kritikal secara bisnis karena mengukur kemampuan model mendeteksi sebanyak mungkin prospek yang sesungguhnya akan melakukan konversi kegagalan mendeteksinya berarti hilangnya peluang pendapatan secara langsung [7]. Peningkatan *F1-Score* sebesar 0,56% juga mengkonfirmasi bahwa keseimbangan antara presisi dan sensitivitas model secara keseluruhan membaik dengan penerapan SMOTE.

Penurunan kecil pada *Precision* sebesar 0,20% merupakan *trade-off* yang wajar dan dapat diantisipasi ketika sensitivitas model terhadap kelas minoritas ditingkatkan melalui teknik *oversampling* [6]. Dalam konteks pemasaran, biaya dari *false negative* (kehilangan prospek potensial yang tidak terdeteksi) jauh lebih besar secara bisnis dibandingkan biaya dari *false positive* (menghubungi prospek yang tidak jadi membeli). Dengan demikian, penurunan *precision* yang sangat kecil ini dinilai dapat diterima sepenuhnya mengingat besarnya keuntungan dari peningkatan deteksi prospek potensial.



Gambar 4. Perbandingan Performa RF Tanpa SMOTE vs RF+SMOTE pada Seluruh Metrik Evaluasi

Ringkasan perbandingan performa seluruh metrik antara kedua model ditampilkan dalam bentuk diagram batang pada Gambar 4. Visualisasi tersebut secara jelas memperlihatkan superioritas model RF+SMOTE pada empat dari lima metrik evaluasi. Konsistensi peningkatan pada metrik *recall*, *F1-Score*, dan *accuracy* secara serentak memvalidasi bahwa integrasi SMOTE tidak hanya meningkatkan sensitivitas model terhadap kelas minoritas, tetapi juga mempertahankan

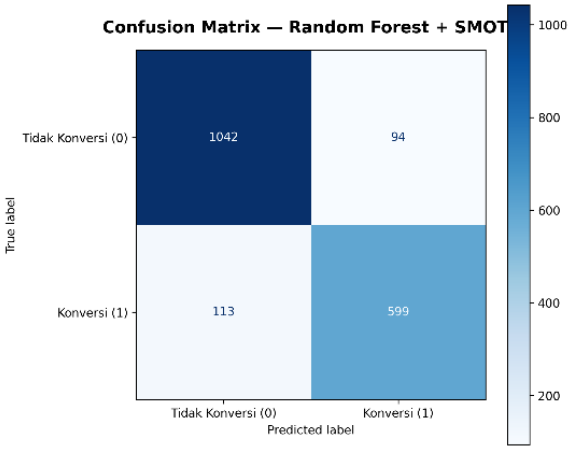
Author : **Daffa Pratama Putra[1)], M. Rizki Al Akbar[2)], Dimas Agil Kusuma[3)], Ali Ibrahim[4)], Fathoni[5)]**

P-ISSN : 2089-3353
E-ISSN : 2808-9162

bahkan sedikit meningkatkan performa keseluruhan model. Temuan ini selaras dengan penelitian [11] yang menyimpulkan bahwa kombinasi teknik penyeimbangan data dengan algoritma *ensemble* menghasilkan performa yang lebih unggul secara komprehensif [11].

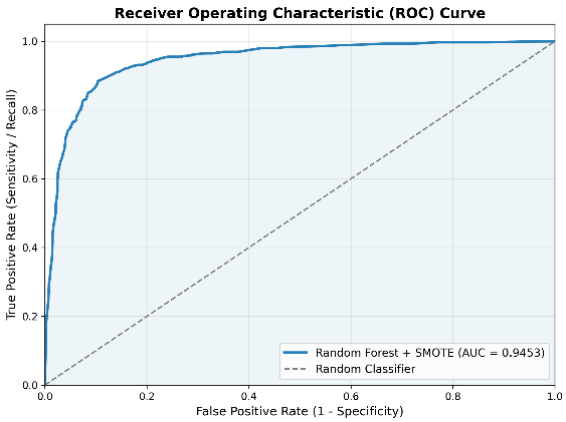
3.4 Analisis *Confusion Matrix* dan Kurva ROC

Hasil klasifikasi model RF+SMOTE pada data uji divisualisasikan melalui *confusion matrix* yang ditampilkan pada Gambar 4. Analisis *confusion matrix* menunjukkan bahwa dari 1.848 sampel data uji, model RF+SMOTE berhasil mengklasifikasikan 1.042 sampel sebagai *True Negative* (TN) dan 599 sampel sebagai *True Positive* (TP), dengan *False Positive* (FP) sebesar 94 dan *False Negative* (FN) sebesar 113. Nilai FN yang relatif kecil mengindikasikan bahwa hanya 113 prospek potensial yang tidak berhasil diidentifikasi oleh model sebuah hasil yang menunjukkan kemampuan deteksi kelas minoritas yang baik pasca penerapan SMOTE. Confusion matrix model random forest+SMOTE ditunjukkan pada Gambar 5 dibawah ini.



Gambar 5. *Confusion Matrix* Model Random Forest + SMOTE

Kemampuan diskriminatif model lebih lanjut dievaluasi menggunakan kurva *Receiver Operating Characteristic* (ROC), sebagaimana ditampilkan pada Gambar 5. Validasi ini menghasilkan nilai AUC sebesar 0,9453, yang tergolong dalam kategori *excellent discrimination* (AUC > 0,90). Kurva ROC yang terbentuk jauh di atas garis *random classifier* pada Gambar 6 mengkonfirmasi bahwa model memiliki kemampuan yang sangat baik dalam membedakan prospek yang berpotensi konversi dari yang tidak, pada berbagai ambang batas keputusan.



Gambar 6. Kurva ROC-AUC Model Random Forest + SMOTE

3.5 Analisis Fitur Paling Berpengaruh

Salah satu keunggulan inheren algoritma Random Forest adalah kemampuannya menghasilkan peringkat kepentingan fitur (*feature importance*) berbasis *Gini Impurity*, yang memberikan *actionable insight* bagi profesional pemasaran dalam mengidentifikasi faktor-faktor paling determinan terhadap konversi. Kelima fitur paling determinan terhadap prediksi konversi beserta nilai *Gini Importance*-nya disajikan secara terurut pada Tabel 2.

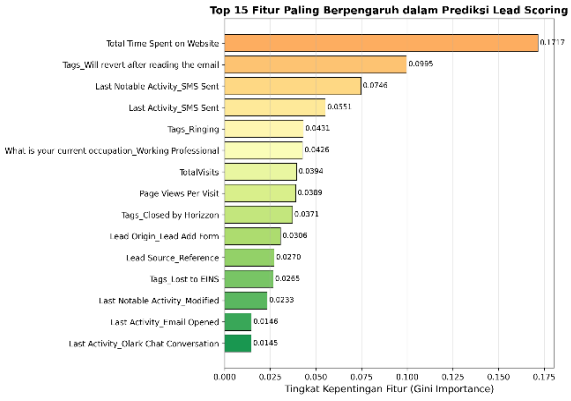
Tabel 2. Lima Fitur Teratas Berdasarkan Gini Importance

Peringkat	Fitur	Gini Importance
1	Total *Time Spent on Website*	0,1717
2	Tags_*Will revert after reading the email*	0,0995
3	*Last Notable Activity_SMS Sent*	0,0746
4	*Last Activity_SMS Sent*	0,0551
5	*Tags_Ringing*	0,0431

Fitur *Total Time Spent on Website* mendominasi sebagai prediktor terkuat *(0,1717) jauh* melampaui fitur lainnya. Temuan ini mengindikasikan bahwa durasi keterlibatan prospek pada platform digital merupakan sinyal konversi yang paling kuat, mencerminkan tingkat minat dan keseriusan yang autentik. Fitur kedua hingga keempat, yaitu *Tags_Will revert after reading the email* (0,0995), *Last Notable Activity_SMS Sent* (0,0746), dan *Last Activity_SMS Sent* (0,0551), secara kolektif menegaskan bahwa responsivitas prospek terhadap aktivitas komunikasi baik melalui email maupun SMS merupakan kelompok sinyal perilaku yang sangat relevan terhadap keputusan konversi. Distribusi kepentingan 15 fitur teratas secara keseluruhan divisualisasikan pada Gambar 6, yang memperkuat dominasi fitur-fitur perilaku digital sebagai prediktor utama konversi. Secara keseluruhan, dominasi fitur-fitur perilaku digital (*behavioral features*) ini selaras dengan argumen [3] bahwa model

P-ISSN : 2089-3353
E-ISSN : 2808-9162

lead scoring berbasis atribut perilaku interaksi menghasilkan akurasi yang lebih superior dibandingkan model berbasis profil demografis semata [3].



Gambar 7. Top 15 Fitur Paling Berpengaruh dalam Prediksi *Lead Scoring*

3.6 Validasi Generalisasi Model

Untuk memastikan hasil tidak bersifat *overfitting* dan dapat digeneralisasi terhadap data baru, validasi dilakukan menggunakan *Stratified 5-Fold Cross-Validation*. Hasil validasi model menggunakan *Stratified 5-Fold Cross-Validation* dirangkum secara lengkap pada Tabel 3, mencakup nilai rata-rata, standar deviasi, dan rentang setiap metrik evaluasi di seluruh *fold*.

Tabel 3. Hasil Stratified 5-*Fold Cross-Validation* Model Random Forest + SMOTE

Metrik	Mean	Std.Dev.	Range
Accuracy	89,76%	± 1,30%	[88,15% – 91,50%]
Precision	88,80%	± 1,35%	[87,29% – 91,23%]
Recall	84,02%	± 2,55%	[80,76% – 86,24%]
F1-Score	86,33%	± 1,86%	[84,05% – 88,66%]
AUC-ROC	**94,91%**	± 0,64%	[94,06% – 95,61%]

Nilai *AUC-ROC* yang sangat konsisten pada rentang 94–95% dengan standar deviasi yang sangat kecil di seluruh *fold* membuktikan bahwa model tidak mengalami *overfitting* terhadap partisi data tertentu. Variansi yang sedikit lebih besar pada metrik *Recall* merupakan karakteristik umum pada metrik berbasis kelas minoritas, namun nilai *recall* yang secara konsisten di atas 80% di seluruh *fold* membuktikan sensitivitas model tetap terjaga dengan baik. Secara keseluruhan, stabilitas hasil *cross-validation* ini mengkonfirmasi bahwa model RF+SMOTE yang diusulkan memiliki kemampuan generalisasi yang tinggi dan andal untuk diimplementasikan dalam sistem pendukung keputusan *lead scoring* pada lingkungan bisnis nyata [11].

4. Kesimpulan

Dalam Penelitian ini berhasil membangun model prediksi *lead scoring* yang tangguh dengan mengintegrasikan algoritma *Random Forest* dan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) untuk mengatasi permasalahan ketidakseimbangan kelas pada dataset pemasaran. Model RF+SMOTE yang diusulkan mencapai nilai *accuracy* 88,80%, *F1-Score* 85,27%, dan *AUC-ROC* 0,9453 melampaui model *baseline* Random Forest tanpa SMOTE pada empat dari lima metrik evaluasi, dengan peningkatan paling signifikan pada metrik *Recall* sebesar 1,26%. Analisis *feature importance* mengungkapkan bahwa fitur perilaku digital, khususnya *Total Time Spent on Website*, merupakan prediktor konversi yang paling dominan, memberikan *insight* praktis bagi profesional pemasaran dalam merancang strategi prioritasi prospek yang lebih efisien. Hasil *5-Fold Cross-Validation* yang stabil dengan nilai *AUC-ROC* konsisten di rentang 94–95% mengkonfirmasi kemampuan generalisasi model yang tinggi. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi variasi teknik *oversampling* yang lebih adaptif seperti SMOTE-ENN atau Borderline-SMOTE, serta menguji implementasi model pada dataset *lead scoring* dari sektor industri yang berbeda guna memperluas generalisasi temuan.

Daftar Rujukan

[1] N. Ahmad, M. J. Awan, H. Nobanee, A. M. Zain, A. Naseem, and A. Mahmoud, "Customer Personality Analysis for Churn Prediction Using Hybrid Ensemble Models and Class Balancing Techniques," *IEEE Access*, vol. 12, pp. 1865–1879, 2024, doi: 10.1109/ACCESS.2023.3334641.

[2] J. Lin, "Application of machine learning in predicting consumer behavior and precision marketing," *PLoS One*, vol. 20, no. 5 May, pp. 1–12, 2025, doi: 10.1371/journal.pone.0321854.

[3] L. González-Flores, J. Rubiano-Moreno, and G. Sosa-Gómez, "The relevance of lead prioritization: a B2B lead scoring model based on machine learning," *Front. Artif. Intell.*, vol. 8, 2025, doi: 10.3389/frai.2025.1554325.

[4] A. Yocupicio-Zazueta, A. Brau-Avila, F. Cirett-Galán, and M. Valenzuela-Galván, "Design and Deployment of ML in CRM to Identify Leads," *Appl. Artif. Intell.*, vol. 38, no. 1, 2024, doi: 10.1080/08839514.2024.2376978.

[5] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00943-4.

[6] M. Altalhan, A. Algarni, and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," *IEEE Access*, vol. 13, pp. 13686–13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[7] A. Manzoor, M. Atif Qureshi, E. Kidney, and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," *IEEE Access*, vol. 12, pp. 70434–70463, 2024, doi: 10.1109/ACCESS.2024.3402092.

[8] E. F. Agyemang *et al.*, "Addressing Class Imbalance Problem in Health Data Classification: Practical Application From an Oversampling Viewpoint," *Appl. Comput. Intell. Soft Comput.*, vol. 2025, no. 1, 2025, doi: 10.1155/acis/1013769.

[9] Z. Zheng, "Financial Risk Early Warning Model Combining SMOTE and Random Forest for Internet Finance Companies," *J. Cases Inf. Technol.*, vol. 26, no. 1, 2024, doi: 10.4018/JCIT.356504.

[10] Husain *et al.*, "SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models," *Algorithms*, vol. 18, no. 1, Jan. 2025, doi: 10.3390/a18010037.

[11] I. Aruleba and Y. Sun, "Effective Credit Risk Prediction Using Ensemble Classifiers With Model Explanation," *IEEE Access*, vol. 12, pp. 115015–115025, 2024, doi: 10.1109/ACCESS.2024.3445308.

[12] B. Amirshahi and S. Lahmiri, "Bankruptcy prediction using optimal ensemble models under balanced and imbalanced data," *Expert Syst.*, vol. 41, no. 8, Aug. 2024, doi: 10.1111/exsy.13599.

[13] S. Gholampour, "Impact of Nature of Medical Data on Machine and Deep Learning for Imbalanced Datasets: Clinical Validity of SMOTE Is Questionable," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 827–841, Jun. 2024, doi: 10.3390/make6020039.

[14] N. S. Thomas and S. Kaliraj, "An Improved and Optimized Random Forest Based Approach to Predict the Software Faults," *SN Comput. Sci.*, vol. 5, no. 5, Jun. 2024, doi: 10.1007/s42979-024-02764-x.

[15] J. Lyu, J. Yang, Z. Su, and Z. Zhu, "LD-SMOTE: A Novel Local Density Estimation-Based Oversampling Method for Imbalanced Datasets," *Symmetry (Basel).*, vol. 17, no. 2, Feb. 2025, doi: 10.3390/sym17020160.