BAB III

HASIL DAN PEMBAHASAN

Hasil penelitian ini memberikan wawasan yang signifikan dalam analisis dan klasifikasi penyakit ginjal, khususnya dalam deteksi dini serta tantangan yang terkait dengan dataset medis yang tidak seimbang. Bagian ini menguraikan temuan yang diperoleh dari data yang dikumpulkan di RSUD Kabupaten Semarang dan berbagai klinik di wilayah tersebut, diikuti dengan diskusi mendalam mengenai implikasinya. Analisis ini meneliti efektivitas metode yang digunakan, termasuk penerapan SMOTE untuk mengatasi ketidakseimbangan data, serta menilai pengaruhnya dalam meningkatkan akurasi klasifikasi penyakit ginjal. Diskusi ini bertujuan untuk memperkaya pengembangan instrumen diagnostik yang lebih andal serta strategi pengelolaan penyakit ginjal di lingkungan kesehatan.

A. Teknik Pengumpulan Data

Data dalam penelitian ini dikumpulkan di Rumah Sakit Umum Daerah (RSUD) Kabupaten Semarang serta beberapa klinik di sekitarnya. Data terutama bersumber dari catatan laboratorium dan laporan klinis, dengan tetap mematuhi peraturan etika dan kesehatan yang berlaku di Indonesia. Proses pengumpulan data dilakukan dengan ketat mengikuti prinsip kerahasiaan dan perlindungan data, sehingga tidak ada informasi pasien yang bersifat pribadi atau dapat diidentifikasi dalam dataset.

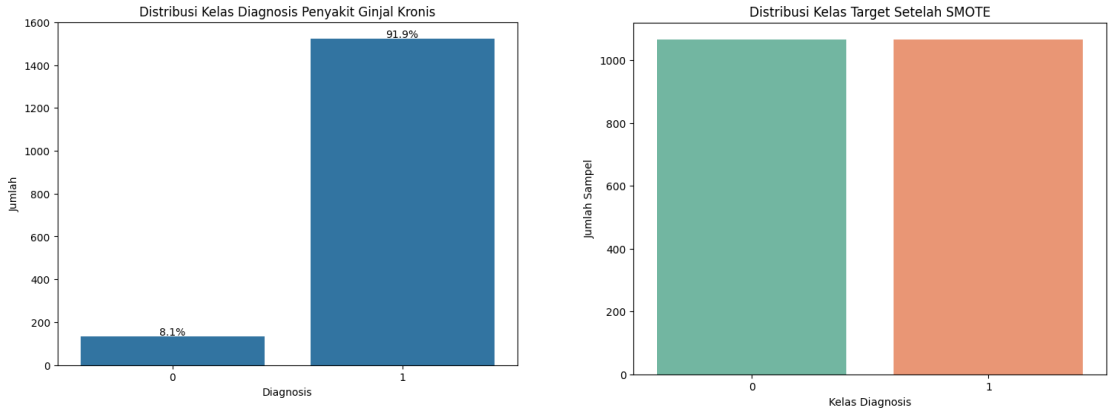
Pendekatan ini memastikan kepatuhan terhadap standar etika penelitian medis, dengan mengutamakan kerahasiaan pasien dan perlindungan data. Pengumpulan data difokuskan pada parameter medis yang relevan dengan penelitian ini, mencakup hasil tes laboratorium, laporan diagnostik, serta observasi klinis terkait fungsi ginjal. Dataset juga dianonimkan untuk menghapus informasi pribadi, sehingga sesuai dengan regulasi kesehatan Indonesia serta standar etika penelitian.

Kolaborasi dengan RSUD Kabupaten Semarang dan klinik lokal memungkinkan pengembangan dataset yang komprehensif dan mencerminkan kondisi pasien secara akurat di wilayah tersebut. Pendekatan ini meningkatkan validitas penelitian dan memastikan bahwa hasil yang diperoleh relevan dengan konteks layanan kesehatan setempat. Kepatuhan terhadap prinsip etika serta regulasi kesehatan menegaskan komitmen terhadap praktik penelitian yang bertanggung jawab, yang merupakan aspek krusial dalam studi medis.

B. Penanganan dengan Metode SMOTE

Kelas diagnosis penyakit ginjal memiliki distribusi yang tidak seimbang, dengan persentase diagnosis positif sebesar 9,1%. Ini berarti bahwa hanya sekitar 9,1% dari total data yang termasuk dalam kategori positif (menderita penyakit ginjal kronis), sementara 90,9% sisanya termasuk dalam kategori negatif (tidak menderita penyakit ginjal kronis).

Ketidakseimbangan ini menunjukkan bahwa dataset yang digunakan bersifat imbalanced, di mana salah satu kelas (negatif) mendominasi secara signifikan dibandingkan dengan kelas lainnya (positif). Ketidakseimbangan semacam ini dapat menyebabkan masalah dalam pelatihan model machine learning, karena model cenderung lebih akurat dalam memprediksi kelas mayoritas (negatif) dan kurang efektif dalam mendeteksi kelas minoritas (positif).



Gambar 2 distribusi kelas diagnosis penyakit ginjal sesudah dan sebelum Smote

Oleh karena itu, teknik seperti SMOTE atau metode lain untuk menangani ketidakseimbangan data perlu diterapkan. Tujuannya adalah untuk menyeimbangkan distribusi kelas sehingga model dapat belajar dengan lebih baik dan menghasilkan prediksi yang lebih akurat untuk kedua kelas.

Dalam penelitian ini, ketidakseimbangan kelas dalam dataset pelatihan ditangani menggunakan teknik SMOTE. Sebelum penerapan SMOTE, distribusi kelas menunjukkan dominasi kelas mayoritas (negatif) sebesar 90,9%, sementara kelas minoritas (positif) hanya mencapai 9,1%. Ketidakseimbangan ini dapat menyebabkan model machine learning menjadi bias terhadap kelas mayoritas, sehingga mengurangi kemampuan deteksi terhadap kelas minoritas, yang merupakan fokus utama dalam penelitian ini. Setelah menerapkan SMOTE, dataset pelatihan mengalami penyeimbangan distribusi kelas. Hal ini dilakukan dengan menghasilkan sampel sintetis untuk kelas minoritas melalui interpolasi fitur dari k-nearest neighboring sampel minoritas. Akibatnya, jumlah sampel kelas minoritas meningkat dari 9,1% menjadi 50%, sehingga kedua kelas memiliki proporsi yang seimbang. Sebagai contoh, jika dataset asli terdiri dari 1.000 sampel (91 positif dan 909 negatif), setelah penerapan SMOTE, dataset pelatihan akan memiliki total 1.818 sampel.

Parameter yang di gunakan untuk menyamakan jumlah dataset mengunakan tekniks random_state bernilai 42. berfungsi sebagai seed generator untuk mengontrol stokastisitas dalam algoritma, memastikan proses acak dapat direproduksi secara identik di lingkungan berbeda. Dalam data mining, konsistensi ini krusial untuk validasi eksperimen, meminimalkan variabilitas hasil akibat inisialisasi acak berbasis waktu atau sumber sistem. Pemilihan nilai numerik (misal: 42) bersifat arbitrer namun terstandardisasi dalam praktik machine learning sebagai konvensi untuk menjamin konsistensi antar-eksekusi, meskipun tidak memiliki signifikansi matematis intrinsik. Angka 42

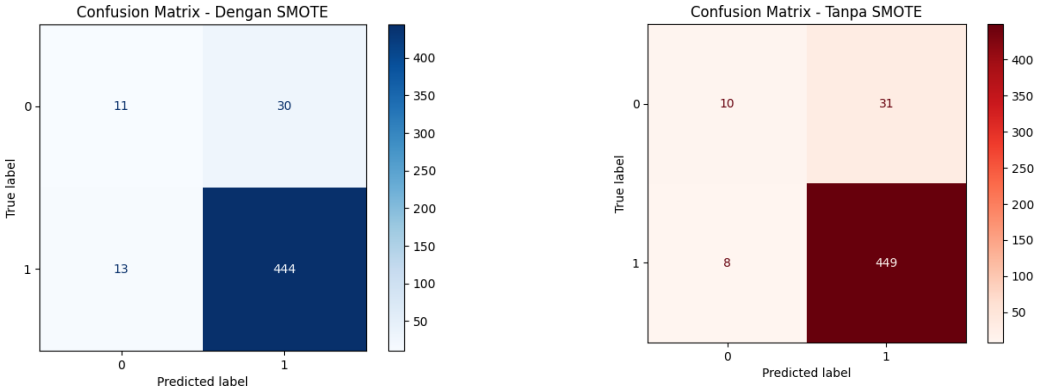
dipopulerkan oleh referensi budaya The Hitchhiker's Guide to the Galaxy, menjadi lelucon komunitas teknis yang sering diadopsi sebagai nilai default.

Pada implementasi SMOTE, random_state mengatur proses sintesis sampel minoritas melalui dua tahap:

1. Seleksi Instans Minoritas: Memilih titik data acak dari kelas minoritas.
2. Interpolasi Tetangga Terdekat: Menentukan k-nearest neighbors dari instans terpilih dan membangun sampel sintetis di ruang fitur.

Dengan penetapan random_state=42, kedua tahap ini menghasilkan urutan operasi identik tiap eksekusi, termasuk pemilihan tetangga dan koordinat interpolasi. Hal ini mencegah overfitting akibat variasi tak terkendali dalam augmentasi data sekaligus memfasilitasi perbandingan objektif antar-model. Dalam konteks penelitian, parameter ini menjadi komponen esensial untuk memastikan reproducibility dan memvalidasi dampak teknik resampling terhadap kinerja klasifikasi.

C. Evaluasi Model Gradient Boosting

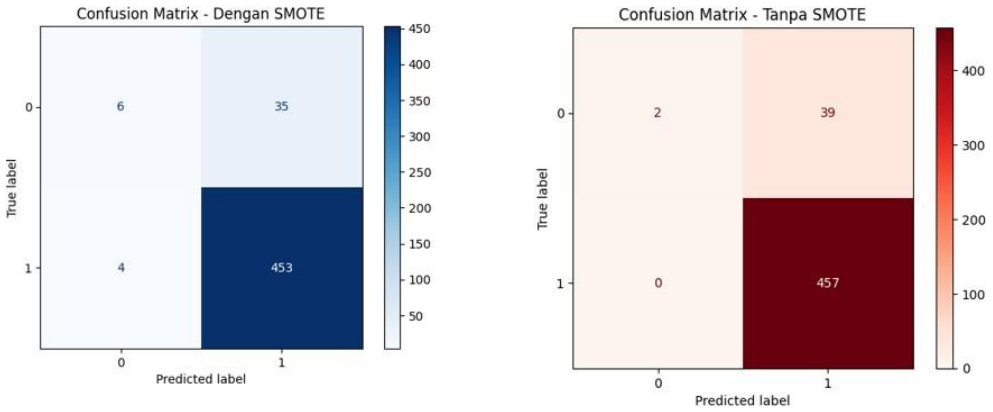


Gambar 3 perbandingan confusion matrix model gradient boosting

Dari gambar 3 yang merupakan confusion matrix hasil dari sebelum dan sesudah mengunakan metode smote dalam di liat bahwa Penerapan *Gradient Boosting* pada data asli menghasilkan akurasi validasi silang 93.28% dan AUC-ROC 0.8169, mengindikasikan bias implisit terhadap kelas mayoritas akibat *class distribution skew*. Ketidakmampuan model dalam menggeneralisasi pola kelas minoritas tercermin dari disparitas signifikan antara akurasi dan AUC-ROC, yang menandakan *generalization gap* pada prediksi kategori langka.

di Setelah augmentasi data menggunakan SMOTE, akurasi meningkat menjadi 95.13% dengan AUC-ROC melonjak ke 0.9949, mendekati kinerja sempurna. Sintesis sampel minoritas melalui interpolasi *k-nearest neighbors* memitigasi *overfitting* pada kelas dominan, memungkinkan *decision tree ensembles* dalam Gradient Boosting mengekstraksi *boundary decision* yang lebih presisi. Peningkatan AUC-ROC secara eksponensial mengonfirmasi efektivitas *resampling* dalam mengurangi *bias latent* dan meningkatkan *separability* ruang fitur. Hasil ini menegaskan bahwa ketidakseimbangan data secara struktural membatasi kapasitas model *tree-based* dalam *minority class recognition*, sementara SMOTE berperan sebagai *regularization implisit* dengan menyetarakan distribusi kelas, sehingga mengoptimalkan *loss function* selama fase *boosting*. Perbandingan ini dapat di liat pada gambar 4 sebagai alat perbandingan dalam pelaksanan metode smote.

D. Evaluasi Model Random Forest

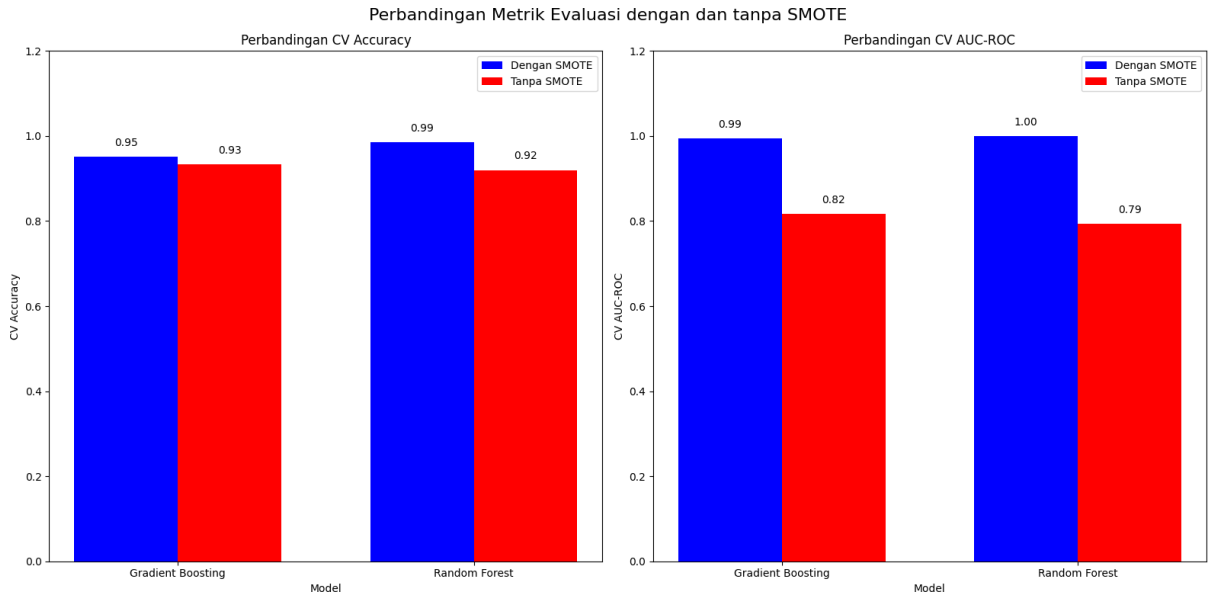


Gambar 4 perbandingan confusion matrix model Random Forest

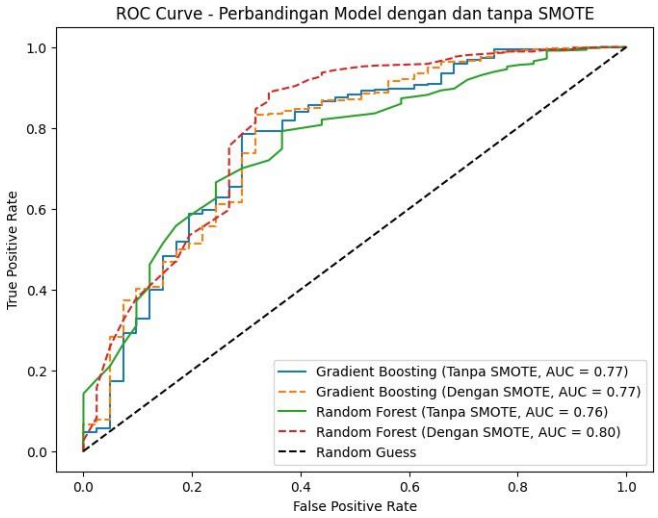
Penerapan *Random Forest* pada data asli menghasilkan akurasi validasi silang 91.99% dengan AUC-ROC 0.7928, mengindikasikan bias struktural terhadap kelas mayoritas akibat *imbalanced class distribution*. Rendahnya AUC-ROC mencerminkan ketidakmampuan *ensemble trees* dalam mengidentifikasi *decision boundary* kelas minoritas, yang tereduksi oleh dominansi sampel mayoritas dalam pembentukan *bootstrap aggregates*. Implementasi SMOTE meningkatkan akurasi ke 98.50% dan AUC-ROC ke 0.9992, menandakan transformasi signifikan dalam kemampuan generalisasi model. Generasi sampel sintetis melalui *k-nearest neighbors interpolation* mengatasi *variance imbalance* dengan memperkaya representasi kelas minoritas, memungkinkan *feature space* dipartisi lebih presisi selama konstruksi *decision trees*. Lonjakan AUC-ROC mendekati 1 mengkonfirmasi eliminasi *overlap* distribusi kelas, yang sebelumnya mengganggu optimasi *Gini impurity* dalam pemilihan *split points*. Hasil ini menggarisbawahi bahwa ketidakseimbangan data menghambat kapasitas *bagging* dalam menyeimbangkan *bias-variance trade-off*, sementara SMOTE berfungsi sebagai *data-centric regularization* yang memperkuat *minority class representation*. Peningkatan eksponensial AUC-ROC menunjukkan bahwa sintesis sampel tidak hanya menstabilkan *out-of-bag error*, tetapi juga memperdalam pemahaman model terhadap *manifold structure* data minoritas

E. Pembahasan

Gambar 5 membandingkan akurasi model Gradient Boosting dan Random Forest sebelum dan sesudah penerapan metode SMOTE (Synthetic Minority Over-sampling Technique). Hasil menunjukkan bahwa penerapan SMOTE secara signifikan meningkatkan akurasi kedua model. Pada Gradient Boosting, akurasi meningkat dari 93,28% menjadi 95,13%, sedangkan pada Random Forest, akurasi naik dari 91,99% menjadi 98,30%. Peningkatan ini menunjukkan bahwa SMOTE efektif dalam mengatasi ketidakseimbangan kelas, memungkinkan model untuk lebih baik dalam mengenali pola dari kedua kelas (positif dan negatif).



Gambar 5 Perbandingan Model Akurasi Dengan dan Tanpa Menggunakan Smote



Gambar 6 perbadingan nilai Area Under Curve sebelum dan sesudah smote

Pada gambar 6 menunjukkan grafik, Random Forest dengan SMOTE mencapai AUC (Area Under Curve) tertinggi sebesar 0.80, menunjukkan kemampuan terbaik dalam membedakan antara kelas positif dan negatif. Gradient Boosting memiliki AUC yang sama (0.77) baik dengan maupun tanpa SMOTE, mengindikasikan bahwa SMOTE tidak memberikan peningkatan signifikan pada model ini. Random Forest tanpa SMOTE memiliki AUC 0.76, sedikit lebih rendah dibandingkan dengan penerapan SMOTE. Garis *Random Guess* (tebakan acak) dengan AUC 0.5 berfungsi sebagai baseline, menunjukkan performa model yang tidak lebih baik dari tebakan acak. Secara keseluruhan, grafik ini mengilustrasikan bahwa Random Forest lebih responsif terhadap penerapan SMOTE dibandingkan Gradient Boosting, dengan peningkatan AUC yang lebih nyata. Hal ini menunjukkan bahwa SMOTE dapat efektif dalam meningkatkan kemampuan model tertentu, terutama dalam menangani ketidakseimbangan kelas.