KLASIFIKASI RISIKO DIABETES MENGGUNAKAN *RANDOM FOREST* DENGAN KOMBINASI SMOTE-TOMEK *LINKS* DAN *RECURSIVE FEATURE ELIMINATION*

SKRIPSI

Oleh :
MUHAMMAD IHYA' ULUMUDDIN AL HAZMI
NIM. 210605110040

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

KLASIFIKASI RISIKO DIABETES MENGGUNAKAN *RANDOM FOREST* DENGAN KOMBINASI SMOTE-TOMEK *LINKS* DAN *RECURSIVE FEATURE ELIMINATION*

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh :
MUHAMMAD IHYA' ULUMUDDIN AL HAZMI
NIM. 210605110040

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

HALAMAN PERSETUJUAN

KLASIFIKASI RISIKO DIABETES MENGGUNAKAN *RANDOM FOREST* DENGAN KOMBINASI SMOTE-TOMEK *LINKS* DAN *RECURSIVE FEATURE ELIMINATION*

SKRIPSI

Oleh :
MUHAMMAD IHYA' ULUMUDDIN AL HAZMI
NIM. 210605110040

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 17 Juni 2025

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono S.Si M.Kom	Dr. Irwan Budi Santoso, M.Kom
NIP. 19680519 200312 1 001	NIP. 19770103 201101 1 004

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

KLASIFIKASI RISIKO DIABETES MENGGUNAKAN *RANDOM FOREST* DENGAN KOMBINASI SMOTE-TOMEK *LINKS* DAN *RECURSIVE FEATURE ELIMINATION*

SKRIPSI

Oleh :
MUHAMMAD IHYA' ULUMUDDIN AL HAZMI
NIM. 210605110040

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 23 Juni 2025

Susunan Dewan Penguji

Ketua Penguji	: Dr. M. Amin Hariyadi, M.T NIP. 19670118 200501 1 001	( )
Anggota Penguji I	: Shoffin Nahwa Utama, M.T. NIP. 19860703 202012 1 003	( )
Anggota Penguji II	: Prof. Dr. Suhartono S.Si M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Muhammad Ihya' Ulumuddin Al Hazmi
NIM	: 210605110040
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Klasifikasi Risiko Diabetes Menggunakan *Random Forest* dengan Kombinasi SMOTE-Tomek *Links* dan *Recursive Feature Elimination*

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 20 Juni 2025
Yang membuat pernyataan,

Muhammad Ihya' Ulumuddin Al Hazmi
NIM.210605110040

MOTTO

"Urip iku Urup"

"Manusia itu semuanya mati(mati perasaannya) kecuali orang–orang yang berilmu. Dan orang yang berilmu itu dalam kebingungan, kecuali mereka yang beramal. Dan mereka yang beramal pun semuanya dalam kekhawatiran kecuali mereka yang ikhlas dan bersih".

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kepada Allah SWT,

Skripsi ini penulis persembahkan untuk:
Kedua orang tuaku Siti Mutawaqqilah dan Suyanto
Tak ada kata-kata yang cukup untuk menggambarkan betapa berharganya kedua orang tua. Semoga anak-anakmu mampu membalas semua kasih sayang, doa dan pengorbananmu dengan kesuksesan di dunia dan akhirat. Aaamiin.

Untuk seluruh saudara dan keluargaku terima kasih atas doa dan dukungannya

Ibu dan Bapak dosen UIN Malang,
Yang telah memberikan bimbingan, dukungan dan ilmu yang sangat berharga sepanjang perjalanan akademik saya.

Sahabat-sahabat saya,
Yang selalu menemani dan memberikan semangat, serta menjadi sumber inspirasi dan pesaing dalam perjalanan akademik.
Semoga kesuksesan selalu bersama kita.

Diri saya sendiri,
Sebagai apresiasi atas usaha, ketekunan dan semangat dalam menyelesaikan penelitian ini, meski penuh tantangan.

KATA PENGANTAR

Alhamdulillah. Segala puji dan syukur hanya milik Allah Subhanahu Wa Ta'ala atas segala nikmat dan kasih sayang-Nya yang telah memudahkan penulis dalam menyelesaikan skripsi ini dengan judul "Klasifikasi Risiko Diabetes Menggunakan *Random Forest* dengan Kombinasi SMOTE-Tomek *Links* dan *Recursive Feature Elimination*". Semoga shalawat dan salam senantiasa tercurah kepada Nabi Muhammad Sallallahu 'Alaihi wa Sallam dan semoga kita semua mendapatkan syafaat beliau di hari kiamat nanti. Aamiin.

Penulis menyampaikan rasa syukur dan terima kasih yang mendalam kepada semua pihak yang telah memberikan dukungan, bantuan, dan motivasi selama proses penyelesaian skripsi ini. Ucapan terima kasih ini ditujukan kepada:

1. Prof. Dr. H. M. Zainuddin, M.A., selaku rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Hariani, M.Si., selaku dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang
3. Dr. Ir. Fachrul Kurniawan, M.MT., IPU selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Suhartono, S.Si., M.Kom., selaku Dosen Pembimbing 1 dan Dr. Irwan Budi Santoso, M.Kom., selaku Dosen Pembimbing 2 yang selalu memberikan arahan saran serta masukan yang sangat berarti dalam proses pengerjaan skripsi.

5. Dr. M. Amin Hariyadi, M.T, selaku Dosen Ketua Penguji dan Shoffin Nahwa Utama, M.T., selaku Dosen Anggota Penguji 1 yang telah menguji serta memberikan masukan sehingga penulis dapat menghasilkan skripsi yang baik.
6. Nia Faricha S, Si., selaku admin Program Studi Teknik Informatika yang selalu sabar memberikan informasi, membantu, dan memberikan arahan selama perkuliahan dan proses penulisan skripsi ini.
7. Seluruh Dosen, Laboran dan Civitas Akademik Program Studi Teknik Informatika yang telah memberikan ilmu pengetahuan, dan dukungan selama studi.
8. Ayah, Ibu, Kakak-Adik dan keluarga besar saya yang selalu memberikan semangat dan doa yang terus dipanjatkan sehingga menjadikan proses penyelesaian skripsi menjadi lebih lancar.
9. Sahabat-sahabat saya dan Semua pihak yang telah memberikan bantuan baik dukungan secara langsung ataupun semangat serta pihak-pihak yang tidak dapat penulis sebutkan.

Penulis menyadari dalam penulisan skripsi ini tidak luput dari kesalahan yang jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun sehingga skripsi ini dapat lebih dikembangkan.

Malang, 20 Juni 2025

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Ulumuddin Al Hazmi, Muhammad Ihya'. 2025. **Klasifikasi Risiko Diabetes Menggunakan *Random Forest* dengan Kombinasi SMOTE-Tomek *Links* dan *Recursive Feature Elimination*** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono S.Si M.Kom, (II) Dr. Irwan Budi Santoso, M.Kom.

.

Kata kunci: Diabetes, Klasifikasi, *Random Forest*, *RFE*, SMOTE-Tomek *Links*.

Penelitian ini bertujuan untuk mengembangkan model klasifikasi risiko diabetes dengan mengimplementasikan algoritma *Random Forest* yang dikombinasikan dengan teknik SMOTE-Tomek Links dan *Recursive Feature Elimination* (RFE), sebagai respons terhadap tingginya prevalensi diabetes dan pentingnya deteksi dini. *Dataset* yang digunakan merupakan *Diabetes Health Indicator Dataset* (DHID), yang memiliki ketidakseimbangan kelas serta kemungkinan adanya fitur yang kurang relevan. Untuk mengatasi permasalahan tersebut, dilakukan empat skenario pengujian: model tanpa penanganan data, model dengan seleksi fitur menggunakan RFE, model dengan penyeimbangan data menggunakan SMOTE-Tomek Links, dan model yang menggabungkan keduanya. Evaluasi dilakukan menggunakan metrik akurasi, presisi, *recall*, dan f1-*score*. Hasil menunjukkan bahwa skenario keempat memberikan performa terbaik, dengan peningkatan signifikan pada *recall* dari 9,01% menjadi 60,49% dan f1-*score* dari 16,14% menjadi 45,05%. Meskipun akurasi dan presisi mengalami sedikit penurunan, model ini jauh lebih efektif dalam mendeteksi pasien yang berisiko diabetes, menunjukkan bahwa pendekatan kombinasi penyeimbangan kelas dan seleksi fitur mampu meningkatkan sensitivitas model dalam konteks klasifikasi medis.

ABSTRACT

Ulumuddin Al Hazmi, Muhammad Ihya‘. 2025. **Diabetes Risk Classification Using Random Forest with SMOTE-Tomek Links and Recursive Feature Elimination**. Thesis. Department of Informatics Engineering, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University of Malang. Supervisors: (I) Prof. Dr. Suhartono S.Si M.Kom, (II) Dr. Irwan Budi Santoso, M.Kom.

Key words: *Classification, Diabetes, Random Forest, RFE, SMOTE-Tomek Links*.

This research aims to develop a diabetes risk classification model by implementing the Random Forest algorithm combined with SMOTE-Tomek Links to handling class imbalance and Recursive Feature Elimination (RFE) to optimize feature selection. This is a response to the high prevalence of diabetes and the critical importance of early detection. The Diabetes Health Indicator Dataset (DHID), which exhibits class imbalance and potentially irrelevant features, was used. To address these issues, four testing scenarios were conducted: a baseline model without data handling, a model with RFE for feature selection, a model using SMOTE-Tomek Links for data balancing, and a comprehensive model combining both techniques. Evaluation was performed using accuracy, precision, recall, and f1-score metrics. The results indicate that the fourth scenario achieved the best performance, showing a significant increase in recall from 9.01% to 60.49% and f1-score from 16.14% to 45.05%. Although accuracy and precision slightly decreased, this combined model proved to be substantially more effective in detecting patients at risk of diabetes, demonstrating that a combined approach of class balancing and feature selection can significantly enhance model sensitivity in the context of medical classification.

البحث مستخلص

علوم الدين الحازمي، محمد إحياء. ٢٠٢٥. **تصنيف خطر الإصابة بمرض السكري باستخدام خوارزمية الغابة العشوائية (Random Forest) مع دمج تقنيت SMOTE-Tomek Links و Recursive Feature Elimination.** رسالة بكالوريوس. برنامج دراسة هندسة المعلومات، كلية العلوم والتكنولوجيا، جامعة الدولة الإسلامية مولانا مالك إبراهيم مالانج. المشرفان: (الأول) الأستاذ الدكتور سوهارتونو، بكالوريوس علوم، ماجستير في علوم الحاسوب. (الثاني) الدكتور إروان بودي سانتوسو، ماجستير في علوم الحاسوب.

الكلمات المفتاحية: السكري، التصنيف، الغابة العشوائية، إزالة الميزات المتكررة ، .SMOTE-Tomek Links

يهدف هذا البحث إلى تطوير نموذج لتصنيف مخاطر مرض السكري من خلال تطبيق خوارزمية الغابة العشوائية (Random Forest)، بالاشتراك مع تقنيتي روابط SMOTE-Tomek لمعالجة عدم توازن الفئات وإزالة الميزات المتكررة (Recursive Feature Elimination - RFE) لاختيار الميزات المثلى. يأتي هذا استجابة للانتشار المتزايد لمرض السكري والأهمية البالغة للكشف المبكر عنه. تم استخدام مجموعة بيانات مؤشر صحة السكري , التي تتميز بعدم توازن الفئات واحتمال وجود ميزات غير ذات صلة. لمعالجة هذه المشكلات، تم إجراء أربعة سيناريوهات اختبار: نموذج أساسي بدون معالجة البيانات، نموذج مع اختيار الميزات باستخدام RFE، نموذج مع موازنة البيانات باستخدام SMOTE-Tomek Links، ونموذج شامل يجمع بين التقنيتين. تم التقييم باستخدام مقاييس الدقة (accuracy)، والضبط (precision)، والاستدعاء (recall)، ودرجة f1-score. تشير النتائج إلى أن السيناريو الرابع حقق أفضل أداء، مع زيادة كبيرة في الاستدعاء من 9.01% إلى 60.49% ودرجة f1-score من 16.14% إلى 45.05%. على الرغم من الانخفاض الطفيف في الدقة والضبط، أثبت هذا النموذج المدمج فعاليته بشكل كبير في الكشف عن المرضى المعرضين لخطر الإصابة بالسكري، مما يدل على أن النهج المشترك لموازنة الفئات واختيار الميزات يمكن أن يعزز بشكل كبير حساسية النموذج في سياق التصنيف الطبي

BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes merupakan salah satu penyakit kronis yang terus meningkat prevalensinya secara global. Menurut *International Diabetes Federation* (IDF) dalam *IDF Diabetes Atlas* edisi ke 10, pada tahun 2021, lebih dari 537 juta orang di seluruh dunia menderita diabetes, dan angka ini diperkirakan akan meningkat menjadi 643 juta pada tahun 2030 dan 783 juta pada tahun 2045. Penyakit ini menjadi ancaman serius terhadap kesehatan masyarakat karena dapat menimbulkan komplikasi seperti penyakit jantung, stroke, kebutaan, gagal ginjal, dan amputasi pada bagian tubuh tertentu (Alva, Gray, Mihaylova, & Clarke, 2014). Selain itu, meningkatnya prevalensi diabetes juga berdampak pada beban ekonomi yang signifikan, baik bagi individu maupun sistem kesehatan di berbagai negara (Rezki, Mazdadi, Indriani, Saragih, & Athavale, 2024).

Diabetes sering berkembang tanpa disadari karena gejalanya yang tidak selalu tampak jelas pada tahap awal. Hal ini menyebabkan banyak orang tidak terdiagnosis hingga kondisi mereka memburuk atau komplikasi serius muncul. Identifikasi risiko diabetes pada tahap awal menjadi sangat penting untuk mencegah perkembangan penyakit dan mengurangi angka morbiditas dan mortalitas (Abdulhadi & Al-Mousa, 2021). Namun, proses deteksi dini sering kali menghadapi tantangan besar, terutama karena ketergantungan pada metode

diagnosis tradisional seperti pemeriksaan fisik dan laboratorium yang memerlukan waktu, biaya, serta akses yang tidak selalu merata (Herman dkk., 2015).

Dalam perspektif Islam, kesehatan dipandang sebagai amanah dari Allah SWT yang harus dijaga dengan baik. Islam tidak hanya mengajarkan pentingnya menjaga kesehatan, tetapi juga menganjurkan umatnya untuk mencari pengobatan apabila mengalami penyakit. Hal ini sesuai dengan sabda Rasulullah ﷺ:

إِنَّ اللَّهَ أَنْزَلَ الدَّاءَ وَالدَّوَاءَ، وَجَعَلَ لِكُلِّ دَاءٍ دَوَاءً فَتَدَاوَوْا وَلَا تَدَاوَوْا بِحَرَامٍ

"Sesungguhnya Allah menurunkan penyakit dan obatnya dan menjadikan bagi setiap penyakit ada obatnya. Maka berobatlah kalian, dan jangan kalian berobat dengan yang haram." (HR. Abu Dawud dari Abu Darda).

Hadis ini menegaskan bahwa setiap penyakit memiliki solusi pengobatan yang telah Allah sediakan. Oleh karena itu, manusia dianjurkan untuk berusaha menemukan cara yang tepat dalam mencegah dan mengobati penyakit, termasuk dalam menghadapi penyakit kronis seperti diabetes. Penyakit ini dapat menyebabkan berbagai komplikasi berbahaya jika tidak terdeteksi dan ditangani sejak dini. Oleh karena itu, penting untuk memiliki sistem yang dapat membantu mendeteksi risiko diabetes secara lebih cepat dan akurat.

Teknologi modern seperti *machine learning* dapat dimanfaatkan sebagai salah satu usaha manusia untuk mendeteksi dan mencegah risiko diabetes sejak dini. *Machine learning* memiliki kemampuan untuk menganalisis data dalam skala besar dan menemukan pola-pola kompleks yang sulit dideteksi oleh metode tradisional. Salah satu algoritma yang banyak digunakan dalam tugas klasifikasi adalah *Random Forest*, yang dikenal karena kemampuannya dalam menghasilkan

prediksi yang akurat serta lebih tahan terhadap *overfitting* dibandingkan metode lain seperti *Decision Tree* (Scornet, Biau, & Vert, 2015). Algoritma ini bekerja dengan membangun sejumlah pohon keputusan acak dan menghasilkan keputusan berdasarkan hasil voting dari pohon-pohon tersebut.

Beberapa penelitian sebelumnya telah memanfaatkan metode *machine learning* untuk membantu dalam deteksi risiko diabetes. Misalnya, penelitian oleh (Chen, 2023) menggunakan *feature engineering* seperti *Pearson Correlation* dan model *Decision Tree* untuk seleksi fitur pada *dataset Diabetes Health Indicator Dataset* (DHID). Meskipun pendekatan ini membantu meningkatkan akurasi prediksi, penelitian tersebut tidak memperhatikan masalah ketidakseimbangan kelas, di mana jumlah pasien yang tidak menderita diabetes jauh lebih besar dibandingkan yang terdiagnosis. Ketidakseimbangan kelas ini mengakibatkan algoritma lebih cenderung memprioritaskan prediksi pada kelas mayoritas, sehingga menurunkan akurasi prediksi pada kelas minoritas (yaitu pasien yang berisiko tinggi diabetes) (Chowdhury, Ayon, & Hossain, 2024).

Selain itu, penelitian oleh (Ullah dkk., 2022) menggunakan metode SMOTE-ENN (*Synthetic Minority Over-sampling Techniquze with Edited Nearest Neighbor*) untuk mengatasi ketidakseimbangan kelas pada *dataset* DHID. Penelitian ini membandingkan penerapan SMOTE-ENN pada berbagai algoritma *machine learning.* Hasil dari penelitian ini menunjukkan evaluasi model penerapan SMOTE-ENN pada algoritma KNN melebihi algoritma lainnya dengan akurasi 98.36%, *precision*, *recall*, dan *f-measures* 98%, serta ROC/AUC *score* 98.3%. Namun, pendekatan ini tidak memperhatikan seleksi fitur, yang sangat penting

untuk memastikan bahwa hanya fitur-fitur yang relevan yang digunakan oleh model. Fitur yang tidak relevan dalam *dataset* dapat menurunkan performa model secara signifikan dan menyebabkan masalah *overfitting* (Zhou, Xin, & Li, 2023).

Berdasarkan hasil tinjauan literatur ini, terdapat gap penelitian yang cukup signifikan, yaitu kurangnya kombinasi antara penanganan ketidakseimbangan kelas dengan seleksi fitur yang optimal. Penelitian sebelumnya lebih banyak fokus pada salah satu aspek saja, baik itu *balancing* kelas atau optimasi seleksi fitur, namun belum banyak yang menggabungkan kedua pendekatan ini secara komprehensif untuk meningkatkan performa model klasifikasi diabetes.

Dalam Islam terdapat anjuran untuk mengerjakan segala sesuatu dengan profesionalisme dan kesungguhan. Rasulullah ﷺ bersabda:

إِنّ اللَّهَ يُحِبّ إِذَا عَمِلَ أَحَدُكُمْ عَمَلاً أَنْ يُتْقِنَهُ

"Sesungguhnya Allah mencintai seseorang yang apabila bekerja, mengerjakannya secara itqan (tepat, terarah, jelas dan tuntas)." (HR. Thabrani dalam al-Mu'jam al-Awsat, No. 897).

Istilah *itqan* sendiri banyak ditemukan dalam Al-Qur'an dan hadis, yang berkaitan dengan amal perbuatan seorang muslim. Rasulullah SA W menganjurkan umatnya untuk selalu beramal dengan sikap *itqan*, yang secara bahasa berarti 'tepat, jelas, dan tuntas' (Hambali & Mualimin, 2021). Konsep ini menekankan pentingnya ketepatan, kesungguhan, dan kecermatan dalam melaksanakan suatu pekerjaan agar menghasilkan *output* yang optimal.

Dalam konteks penelitian ini, konsep *itqan* tercermin dalam upaya mengoptimalkan proses analisis data dan klasifikasi risiko diabetes dengan

memanfaatkan algoritma *Random Forest*, serta teknik optimasi kombinasi SMOTE-Tomek *Links* dan seleksi fitur *Recursive Feature Elimination* (RFE), penelitian ini bertujuan untuk menghasilkan model yang lebih akurat dan efisien dalam mendeteksi risiko diabetes. Hal ini selaras dengan ajaran Islam yang mengajarkan bahwa pekerjaan yang dilakukan dengan kesungguhan dan ketelitian akan membawa manfaat yang lebih besar bagi masyarakat. Dengan demikian, penelitian ini diharapkan dapat berkontribusi dalam pengembangan teknologi kesehatan yang lebih baik dan bermanfaat bagi umat manusia.

1.2 Rumusan Masalah

Bagaimana mengimplementasikan algoritma *Random Forest* dengan kombinasi SMOTE-Tomek *Links* dan *Recursive Feature Elimination* (RFE) serta menganalisis performa model dalam klasifikasi risiko diabetes?

1.3 Batasan Masalah

Untuk memfokuskan penelitian ini, beberapa batasan masalah diterapkan sebagai berikut:

1. Penelitian ini hanya akan menggunakan algoritma *Random Forest* untuk klasifikasi risiko diabetes.
2. *Dataset* yang digunakan yaitu *Diabetes Health Indicators Dataset*.
3. Metode SMOTE-TOMEK *Links* digunakan untuk mengatasi ketidakseimbangan data/kelas.

4. Proses *feature selection* dilakukan dengan menggunakan metode *Recursive Feature Elimination* (RFE) untuk memilih fitur yang paling berkontribusi terhadap akurasi model.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah mengimplementasikan algoritma *Random Forest* dengan kombinasi SMOTE-TOMEK *Links* dan *Recursive Feature Elimination (RFE)* serta menganalisis performa model dalam klasifikasi risiko diabetes.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Bagi pengembangan ilmu pengetahuan, penelitian ini berkontribusi pada pengembangan metode klasifikasi berbasis *machine learning*, khususnya dalam penanganan ketidakseimbangan kelas pada *dataset* kesehatan menggunakan kombinasi metode SMOTE-TOMEK *Links* dan seleksi fitur dengan RFE.

2. Bagi masyarakat dan bidang kesehatan, penelitian ini dapat membantu dalam deteksi dini risiko diabetes, sehingga dapat menurunkan angka morbiditas dan mortalitas akibat penyakit ini.

3. Bagi praktisi *machine learning*, penelitian ini dapat memberikan wawasan tentang bagaimana memanfaatkan metode *Random Forest* dengan optimasi SMOTE-TOMEK dan RFE untuk meningkatkan kinerja prediksi pada *dataset* dengan ketidakseimbangan kelas.

4. Bagi pengembangan sistem pendukung keputusan di bidang kesehatan, hasil penelitian ini dapat menjadi dasar dalam pengembangan aplikasi atau sistem prediksi risiko diabetes yang lebih akurat dan efisien

BAB II
STUDI PUSTAKA

2.1 Penelitian Terkait

Penelitian yang dilakukan (Chen, 2023) bertujuan untuk memprediksi risiko diabetes menggunakan *Diabetes Health Indicators Dataset* dari Kaggle. Chen membandingkan lima algoritma *machine learning*, yaitu *Random Forest*, *Logistic Regression*, *Decision Tree*, *Support Vector Machine (SVM)*, dan *XGBoost*, serta *Deep Neural Network* (DNN). Penelitian ini juga menggunakan *Pearson Correlation* dan *Decision Tree* untuk seleksi fitur pada *dataset*. Hasil penelitian menunjukkan bahwa model DNN menghasilkan akurasi prediksi tertinggi sebesar 75.49% dengan tingkat *recall* mencapai 79%. Indikator evaluasi yang digunakan dalam penelitian ini adalah akurasi, AUC, dan *recall*, yang semuanya diukur untuk membandingkan kinerja dari setiap model yang diuji.

Pada penelitian yang dilakukan (Ullah dkk., 2022) mengeksplorasi teknik SMOTE-ENN (*Synthetic Minority Over-sampling Technique with Edited Nearest Neighbor*) untuk mengatasi ketidakseimbangan kelas pada *dataset Diabetes Health Indicators Dataset* dari Kaggle. Dalam penelitian ini, beberapa algoritma *machine learning* diterapkan, seperti *Random Forest*, K-Nearest Neighbors (KNN), XGBoost, AdaBoost, dan Bagging. Hasil penelitian menunjukkan bahwa penerapan SMOTE-ENN pada algoritma KNN menghasilkan akurasi tertinggi, yaitu sebesar 98,36%, dengan nilai *precision*, *recall*, *f-measure*, dan AUC yang juga tinggi. Penelitian ini menekankan pentingnya penggunaan metode penyeimbangan data untuk meningkatkan kinerja algoritma klasifikasi

Selanjutnya, penelitian yang dilakukan (Shabrina Assyifa & Luthfiarta, 2024) membahas penggunaan teknik *re-sampling* untuk mengatasi ketidakseimbangan data dalam klasifikasi multikelas, Penelitian ini menggunakan algoritma *Random Forest* sebagai model utama dan menerapkan teknik SMOTE-Tomek untuk menyeimbangkan data. Selain itu, optimalisasi parameter model dilakukan menggunakan GridSearchCV. Hasil penelitian menunjukkan adanya peningkatan akurasi model dari 84% menjadi 85%, serta peningkatan nilai *precision*, *recall*, dan F1-Score masing-masing sebesar 85%. Hal ini menunjukkan bahwa teknik SMOTE-Tomek mampu meningkatkan akurasi model klasifikasi, khususnya dalam menangani masalah ketidakseimbangan data.

Penelitian yang dilakukan (Dhani, Siswa, & Pranoto, 2024) bertujuan untuk meningkatkan akurasi klasifikasi stunting dengan memanfaatkan metode *Random Forest* yang dikombinasikan dengan seleksi fitur menggunakan ANOVA dan teknik SMOTE untuk mengatasi ketidakseimbangan kelas dalam data. *Dataset* yang digunakan dalam penelitian ini diambil dari Dinas Kesehatan Kota Samarinda, dengan jumlah 150.466 data dan 21 atribut. Validasi model dilakukan menggunakan k-fold cross-validation dengan k=10. Hasil penelitian menunjukkan bahwa akurasi model meningkat dari 98,83% menjadi 99,77% setelah penerapan seleksi fitur ANOVA dan teknik SMOTE. Penelitian ini menunjukkan bahwa penggabungan metode seleksi fitur dan penyeimbangan data mampu meningkatkan kinerja model.

Pada penelitian yang dilakukan (Wang dkk., 2021) mengeksplorasi berbagai metode *supervised classifiers* untuk prediksi Diabetes Mellitus (DM) guna

mengatasi masalah dimensi fitur tinggi dan ketidakseimbangan data. *Dataset* yang digunakan merupakan data survei yang dilakukan peneliti dengan jumlah 4105 data. Penelitian ini menghadapi tantangan berupa dimensi fitur yang tinggi dan ketidakseimbangan data. Dalam penelitian ini, digunakan kombinasi teknik SVM-SMOTE dan dua metode seleksi fitur, yaitu *Logistic Stepwise Regression* dan LASSO. Hasil penelitian menunjukkan bahwa model *Random Forest* dengan kombinasi SVM-SMOTE dan LASSO menghasilkan kinerja terbaik, dengan akurasi sebesar 89%, *precision* 86,9%, *recall* 91,9%, F1-Score 89,3%, dan AUC 94,8%. Penelitian ini menunjukkan bahwa kombinasi teknik penanganan ketidakseimbangan data dan seleksi fitur mampu meningkatkan kinerja model klasifikasi diabetes secara signifikan.

Tabel 2. 1 Rangkuman Penelitian Terkait

No.	Referensi	*Dataset*	Algoritma/Metode yang Digunakan	Hasil
1	(Chen, 2023)	*Diabetes Health Indicators Dataset*	Menggunakan Pearson Correlation dan Decision Tree untuk seleksi fitur. Algoritma yang diuji: *Random Forest*, Logistic Regression, SVM, XGBoost, DNN.	Model DNN menghasilkan akurasi 75.49% dan recall 79%.
2	(Ullah dkk., 2022)	*Diabetes Health Indicators Dataset*	Teknik SMOTE-ENN untuk penyeimbangan data dan algoritma: *Random Forest*, KNN, XGBoost, AdaBoost, Bagging.	KNN dengan SMOTE-ENN menghasilkan akurasi tertinggi sebesar 98,36%.
3	(Shabrina Assyifa & Luthfiarta, 2024)	Data ulasan aplikasi seluler	*Random Forest* sebagai model utama, teknik SMOTE-Tomek untuk penyeimbangan data, dan GridSearchCV untuk optimasi parameter.	Peningkatan akurasi model dari 84% menjadi 85%, *precision*, recall, dan F1-Score juga meningkat menjadi 85%.
4	(Dhani dkk., 2024)	Data dari Dinas Kesehatan Kota Samarinda	*Random Forest* dikombinasikan dengan seleksi fitur ANOVA dan teknik SMOTE.	Peningkatan akurasi model dari 98,83% menjadi 99,77%

No.	Referensi	*Dataset*	Algoritma/Metode yang Digunakan	Hasil
5	(Wang dkk., 2021)	Data survei dengan 4105 entri	Kombinasi SVM-SMOTE untuk penyeimbangan data, Logistic Stepwise Regression dan LASSO untuk seleksi fitur.	*Random Forest* dengan kombinasi SVM-SMOTE dan LASSO menghasilkan akurasi 89%, *precision* 86,9%, recall 91,9%, F1-Score 89,3%, dan AUC 94,8%.
6	Peneliti	*Diabetes Health Indicators Dataset*	Implementasi *Random Forest* dengan Kombinasi SMOTE-TOMEK *Links* untuk penyeimbangan data dan RFE untuk seleksi fitur.	Hasil yang diharapkan adalah model dapat menghasilkan kemampuan prediksi yang baik.

Dari berbagai penelitian yang telah dilakukan mengenai klasifikasi risiko diabetes, penanganan ketidakseimbangan data dan seleksi fitur, tidak ada yang secara komprehensif menggabungkan *Random Forest* dengan SMOTE-Tomek *Links* untuk menangani ketidakseimbangan data serta RFE untuk seleksi fitur dalam klasifikasi risiko diabetes. Penelitian ini relevan karena mengisi gap tersebut dengan menawarkan solusi yang lebih komprehensif dan optimal dalam meningkatkan akurasi prediksi risiko diabetes. Integrasi antara SMOTE-Tomek *Links* untuk menangani ketidakseimbangan data dan RFE untuk seleksi fitur merupakan pendekatan yang belum dieksplorasi secara penuh dalam penelitian-penelitian terdahulu.

2.2 Diabetes Mellitus

Diabetes mellitus (DM) merupakan penyakit kronis yang ditandai dengan meningkatnya kadar gula darah akibat gangguan metabolisme pada pankreas. Diabetes melitus terbagi menjadi dua tipe, yaitu diabetes tipe 1 dan diabetes tipe 2. Diabetes tipe 1 disebabkan oleh penyakit *autoimun* yang merusak sel-sel β

pankreas, sementara diabetes tipe 2 disebabkan oleh kombinasi faktor genetik dan lingkungan, seperti obesitas, pola makan, kurang aktivitas fisik, stres, dan penuaan (Ozougwu, 2013). Pada diabetes tipe 2, terdapat resistensi insulin serta gangguan pada sekresi insulin yang menyebabkan hiperglikemia (peningkatan kadar gula darah) (Dewi, Azizah, Rendowaty, Sri Wahyuni, & Pranata, 2023).

Menurut *International Diabetes Federation* (IDF), pada tahun 2021, lebih dari 537 juta orang di seluruh dunia menderita diabetes, dan angka ini diperkirakan akan meningkat menjadi 643 juta pada tahun 2030 dan 783 juta pada tahun 2045. Penyakit ini menjadi ancaman serius terhadap kesehatan masyarakat karena dapat menimbulkan komplikasi seperti kerusakan ginjal, kerusakan saraf, penyakit jantung, stroke, hingga kematian. Diabetes juga merupakan salah satu 10 penyebab kematian teratas di seluruh dunia (Ismail, Materwala, & Al Kaabi, 2021). Deteksi dini dan manajemen risiko diabetes menjadi sangat penting dalam upaya pencegahan dan pengendalian penyakit ini. Sehingga bisa dilakukan intervensi medis lebih awal dan membantu menurunkan angka kejadian penyakit diabetes dan komplikasinya.

Adapun langkah pencegahan untuk penyakit ini dapat dilakukan dengan menjaga pola makan sehat, olahraga teratur, menjaga berat badan ideal, manajemen stress, dan pemeriksaan gula darah rutin. Selain itu menghindari gaya hidup tidak sehat seperti kebiasaan merokok dan konsumsi alkohol juga dapat mengurangi risiko terkena diabetes (Suratun, Pujiana, & Sari, 2023).

2.3 Klasifikasi

Klasifikasi merupakan salah satu tugas fundamental dalam *machine learning* yang bertujuan untuk memprediksi label atau kategori dari suatu kelas berdasarkan karakteristik atau fitur-fiturnya (Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, & Susanti, 2022). Dalam konteks *supervised learning*, model klasifikasi dilatih menggunakan *dataset* yang terdiri dari kelas-kelas dengan label yang sudah diketahui, dengan tujuan untuk membangun model yang dapat memprediksi label dari kelas baru yang belum pernah dilihat sebelumnya (Castelli, Vanneschi, & Largo, 2019).

Proses klasifikasi umumnya terdiri dari beberapa tahap utama: *preprocessing* data, ekstraksi dan seleksi fitur, pemilihan dan pelatihan model, serta evaluasi dan optimasi model (Mulyani, Muhamad, & Cahyanto, 2021). *Preprocessing data* melibatkan pembersihan data, penanganan *missing value*, dan normalisasi. Ekstraksi dan seleksi fitur bertujuan untuk mengidentifikasi dan memilih karakteristik yang paling informatif dan relevan untuk tugas klasifikasi. Tahap pemilihan dan pelatihan model melibatkan pemilihan algoritma klasifikasi yang sesuai dan pengoptimalan parameter modelnya. Terakhir, evaluasi model dilakukan untuk menilai performa klasifikasi menggunakan berbagai metrik seperti akurasi, *precision*, *recall*, dan F1-*score* (Tharwat, 2021).

Pemanfaatan klasifikasi *machine learning* telah meluas ke berbagai bidang, termasuk pengenalan gambar, pemrosesan bahasa alami, diagnostik medis, deteksi *fraud*, dan lain-lain. Dalam konteks medis, klasifikasi telah digunakan secara luas

untuk berbagai tugas diagnostik seperti klasifikasi risiko penyakit diabetes, kanker, dan penyakit lainnya (Roihan, Sunarya, & Rafika, 2020).

2.4 *Imbalanced Data*

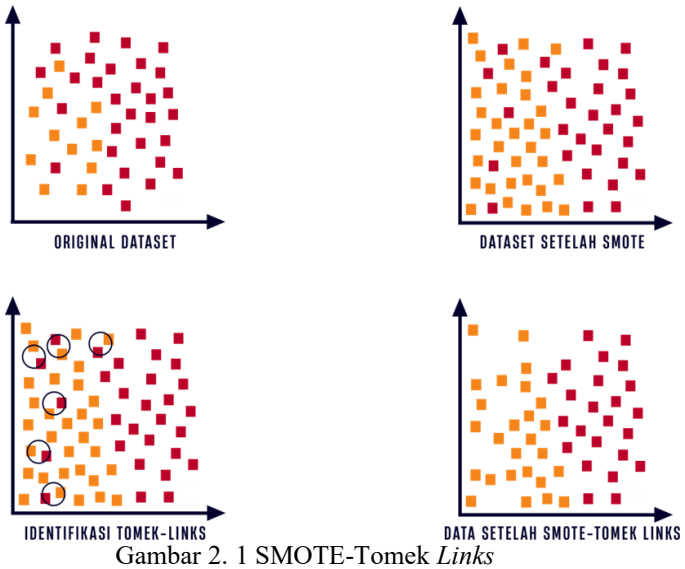
Imbalanced data atau ketidakseimbangan data merupakan kondisi di mana distribusi kelas dalam *dataset* tidak seimbang atau tidak merata (Patel dkk., 2020). Suatu *dataset* dikatakan mengalami ketidakseimbangan ketika jumlah sampel pada satu kelas jauh lebih banyak (mayoritas) dibandingkan dengan kelas lainnya (minoritas). Dalam konteks klasifikasi medis seperti diabetes, sering kali jumlah pasien yang tidak memiliki diabetes (kelas mayoritas) jauh lebih banyak dibandingkan dengan pasien yang memiliki diabetes (kelas minoritas). Data yang tidak seimbang menyebabkan banyak dampak buruk terhadap kinerja model, seperti model cenderung lebih fokus terhadap kelas mayoritas dan mengabaikan kelas minoritas yang menyebabkan performa prediksi terhadap kelas minoritas menjadi buruk dan menurunkan akurasi keseluruhan model (Kumar, Bhatnagar, Gaur, & Bhatnagar, 2021).

Ada beberapa teknik untuk mengatasi ketidakseimbangan data, seperti teknik *under-sampling* dan *over-sampling*. *Under-sampling* menyeimbangkan data dengan cara mengurangi jumlah sampel kelas mayoritas. Sebaliknya *over-sampling* menyeimbangkan data dengan cara menambah jumlah sampel kelas minoritas. Selain itu, ada juga teknik gabungan seperti SMOTE-Tomek *Links*, yang mengkombinasikan kedua metode tersebut.

2.5 SMOTE-TOMEK *Links*

SMOTE-Tomek *Links* merupakan teknik gabungan dari metode SMOTE *over-sampling* dan Tomek *Links under-sampling* yang digunakan untuk menangani masalah ketidakseimbangan kelas (*imbalanced class*) dalam *dataset* (Sir & Soepranoto, 2022). Teknik ini pertama kali diperkenalkan sebagai solusi untuk meningkatkan kinerja klasifikasi pada *dataset* yang tidak seimbang dengan cara menyeimbangkan distribusi kelas mayoritas dan minoritas sekaligus membersihkan *noise* dan *overlap* antar kelas.

SMOTE berfungsi dengan membuat sampel sintetis dari kelas minoritas untuk meningkatkan representasi data pada kelas tersebut (Shabrina Assyifa & Luthfiarta, 2024). Sementara itu, Tomek *Links* digunakan untuk membersihkan *noise* dan *overlap* antar kelas dengan menghapus contoh yang berada pada perbatasan antara kelas mayoritas dan minoritas, yang sering kali menyebabkan kebingungan pada model (Swana, Doorsamy, & Bokoro, 2022).



Gambar 2. 1 SMOTE-Tomek *Links*

Gambar 2.1 merupakan ilustrasi bagaimana SMOTE-Tomek *Links* bekerja memproses data. SMOTE-Tomek *Links* bekerja melalui proses dua tahap yang berurutan untuk menangani ketidakseimbangan data dan membersihkan *noise*. Pada tahap pertama, algoritma SMOTE menganalisis kelas minoritas dalam *dataset* dan mengidentifikasi tetangga terdekat untuk setiap sampel dalam kelas minoritas tersebut. Setelah mengidentifikasi tetangga terdekat, SMOTE menciptakan sampel sintetis baru dengan cara melakukan interpolasi antara sampel asli dan tetangga terpilihnya. Proses interpolasi ini dilakukan dengan mengambil vektor perbedaan antara sampel asli dan tetangganya, kemudian mengalikan vektor tersebut dengan angka acak antara 0 dan 1, dan akhirnya menambahkan hasilnya ke sampel asli. Proses ini diulang hingga jumlah sampel kelas minoritas mencapai keseimbangan dengan kelas mayoritas.

Setelah proses SMOTE selesai, tahap kedua dimulai dengan algoritma Tomek *Links*. Pada tahap ini, algoritma mencari pasangan sampel yang membentuk Tomek *Links*, yaitu pasangan sampel dari kelas berbeda yang menjadi tetangga terdekat satu sama lain dan tidak ada sampel lain yang jaraknya lebih dekat ke salah satu dari pasangan tersebut. Ketika Tomek *Links* ditemukan, sampel dari kelas mayoritas dalam pasangan tersebut dihapus untuk membersihkan *overlap* antara kelas. Proses ini membantu memperjelas batas keputusan antara kelas-kelas yang ada, mengurangi bias dalam klasifikasi, dan meningkatkan kualitas *dataset* secara keseluruhan.

2.6 *Recursive Feature Elimination*

Recursive Feature Elimination (RFE) merupakan metode seleksi fitur berbasis *wrapper* yang bekerja secara rekursif untuk mengeliminasi fitur-fitur yang kurang penting dalam proses prediksi hingga fitur terbaik didapatkan (Pratama, Latipah, & Sari, 2022). RFE bekerja dengan model estimator sebagai komponen utamanya, di mana estimator tersebut harus memiliki kemampuan untuk memberikan peringkat kepentingan fitur, seperti yang dimiliki oleh algoritma *Random Forest* melalui *feature importance scores*.

Proses seleksi fitur menggunakan RFE dilakukan secara iteratif dengan mengeliminasi fitur yang memiliki kepentingan terendah pada setiap iterasi (Satria, Siswa, & Pranoto, 2024). Pada setiap tahap eliminasi, model dilatih ulang menggunakan fitur-fitur yang tersisa, kemudian performa model dievaluasi untuk menentukan kombinasi fitur optimal. RFE juga dapat dikombinasikan dengan *cross-validation* (RFECV) untuk menentukan jumlah fitur optimal secara otomatis, performa model pada setiap *subset* fitur dievaluasi menggunakan validasi silang (Devia & Soewito, 2023).

RFE memiliki keunggulan dalam kemampuannya untuk mempertimbangkan interaksi antar fitur, karena evaluasi dilakukan terhadap *subset* fitur secara keseluruhan, bukan terhadap fitur individual. Hal ini membuat RFE lebih efektif dibandingkan metode seleksi fitur *univariate* yang mengevaluasi fitur secara terpisah.

2.7 *GridSearchCV*

GridSearchCV merupakan salah satu metode *hyperparameter tuning* yang digunakan untuk menemukan kombinasi parameter terbaik pada model *machine learning* dengan cara melakukan pencarian secara menyeluruh di ruang parameter yang telah ditentukan (Rasheed, Kiran Kumar, Rani, Prasad Kantipudi, & M, 2024). *Hyperparameter tuning* bertujuan untuk meningkatkan performa model dengan menentukan nilai *hyperparameter* optimal yang dapat menghasilkan akurasi, precision, recall, dan F1-*score* yang lebih baik.

GridSearchCV bekerja dengan membangun dan mengevaluasi model *machine learning* untuk setiap kombinasi parameter yang mungkin. Selanjutnya, model terbaik dipilih berdasarkan performa metrik evaluasi tertentu, seperti akurasi atau F1-score. Metode ini menggunakan teknik validasi silang (*cross-validation*) untuk memastikan bahwa hasil *tuning* tidak *overfitting* dan tetap *generalizable* ke data baru.

2.8 *Random Forest*

Random Forest merupakan algoritma *ensemble learning* yang dikembangkan oleh Leo Breiman pada tahun 2001, sebagai pengembangan dari metode *bagging* dan *random decision forests* (Ibrahim, 2023). *Random Forest* juga merupakan kumpulan dari beberapa *decision tree* yang masing-masing menghasilkan prediksi, dan hasil akhir diperoleh berdasarkan keputusan terbanyak dari pohon-pohon tersebut. Menurut (Iskandar, Gutama, Wijaya, & Danianti, 2024), algoritma *Random Forest* bekerja melalui beberapa tahapan berikut:

1. *Dataset* asli dibagi menjadi beberapa subset data melalui teknik *Bootstrap Aggregating (Bagging)*, yaitu pengambilan sampel acak dengan pengembalian. Setiap subset digunakan untuk melatih satu pohon keputusan.
2. Dalam proses pembangunan pohon, hanya sebagian fitur yang dipilih secara acak di setiap *node*. Ini dilakukan untuk meningkatkan keragaman antar pohon. Pemisahan data di setiap *node* didasarkan pada nilai Gini Impurity, dengan rumus:

$$Gini(s_i) = 1 - \sum_{i=1}^{c} p_i^2 \quad (2.1)$$

Keterangan :
p_i : proporsi kelas ke-i dalam node
c : jumlah total kelas

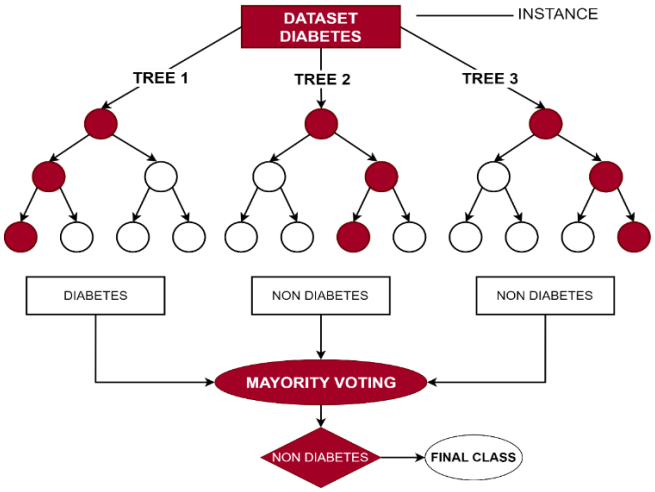
3. Titik *split* terbaik ditentukan berdasarkan nilai Gini *Split,* yaitu nilai rata-rata tertimbang dari impurity dua subset hasil *split:*

$$Gini_{split} = \sum_{i=1}^{k} (\frac{n_i}{n}) \, Gini(s_i) \quad (2.2)$$

Keterangan :
n_i : jumlah data pada *subset* ke-i
n : total data pada *node*
$Gini(s_i)$: nilai gini dari subset ke-i

4. Proses ini diulang secara rekursif untuk setiap node hingga pohon mencapai kedalaman maksimum atau node tidak dapat di-*split* lebih lanjut.
5. Setelah semua pohon terbentuk, prediksi terhadap data uji dilakukan dengan mengambil suara dari masing-masing pohon, lalu hasil akhir

ditentukan berdasarkan kelas yang paling banyak dipilih (mayoritas).



Gambar 2. 2 Alur kerja algoritma *Random Forest*

Sebagai contoh, gambar 2.2 menjelaskan alur kerja algoritma *Random Forest* secara visual pada kasus klasifikasi diabetes. Pada gambar tersebut, dataset diabetes dibagi menjadi tiga subset melalui proses bootstrap. Setiap subset digunakan untuk membentuk satu *decision tree* (*Tree 1, Tree 2, Tree 3*). Pohon-pohon tersebut dilatih secara independen dengan fitur acak yang dipilih pada setiap node, sehingga menghasilkan klasifikasi akhir masing-masing: Tree 1 memprediksi *diabetes*, sementara *Tree 2* dan *Tree 3* memprediksi *non diabetes*. Berdasarkan mekanisme voting mayoritas, dua dari tiga pohon memprediksi *non diabetes*, sehingga hasil akhir klasifikasi untuk instance tersebut adalah *non diabetes*.

Random Forest banyak digunakan untuk tugas klasifikasi karena memiliki beberapa keunggulan yang tidak dimiliki algoritma lain, yaitu kemampuannya untuk mengatasi *overfitting*, toleransi terhadap data yang tidak seimbang, mengatasi *missing value* pada data, serta memiliki kemampuan untuk mengidentifikasi fitur

paling penting *(feature importance)* (Scornet dkk., 2015). Meskipun memiliki banyak keunggulan, penggunaan *Random Forest* dalam klasifikasi juga memiliki beberapa tantangan, seperti bias terhadap data yang tidak seimbang, bias terhadap fitur *categorical* dengan banyak level, serta seleksi fitur yang tidak otomatis.

2.9 *Confusion Matrix*

Evaluasi model merupakan tahapan penting dalam pengembangan model klasifikasi untuk mengukur seberapa baik model dapat melakukan prediksi. Dalam konteks klasifikasi, pemilihan metrik evaluasi yang tepat sangat penting untuk mengukur performa model secara akurat. Salah satu metode yang banyak digunakan untuk mengevaluasi kinerja model klasifikasi adalah *confusion matrix* (Tharwat, 2021). Metode ini mampu memberikan gambaran tentang performa dari model dengan menunjukkan perhitungan hasil prediksi benar dan salah untuk setiap kelas (Awad & Khanna, 2015).

Tabel 2. 2 Struktur *confusion matrix*

Actual	**Predict Diabetic (1)**	**Predict Non-Diabetic (0)**
Diabetic (1)	True Positive	False Negative
Non-Diabetic (0)	False Positive	True Negative

Pada tabel 2.1 struktur *confusion matrix* memiliki empat kategori yaitu *True Positive* (TP) menunjukkan jumlah data *diabetic* yang berhasil diidentifikasi model ke dalam kelas *diabetic*. *False Positive* (FP) menunjukkan jumlah kesalahan model dalam mengidentifikasi data *non-diabetic* ke dalam kelas *diabetic*. *False Negative* (FN) menunjukkan jumlah kesalahan model dalam mengidentifikasi data *diabetic* ke dalam kelas *non-diabetic*. *True Negative* (TN) menunjukkan jumlah data *non-diabetic* yang berhasil diidentifikasi model ke dalam kelas *non-diabetic.* Kategori-

kategori *confusion matrix* tersebut dapat digunakan untuk menghitung berbagai metrik evaluasi, seperti akurasi, *precision*, *recall,* dan *f1-score*.

Akurasi merupakan metrik yang menghitung seberapa banyak model melakukan prediksi dengan benar. Akurasi dapat dihitung dengan cara membagi jumlah prediksi benar dengan total prediksi yang dilakukan.

$$Akurasi = \frac{TP + TN}{TP+TN + FN +FP} \tag{2.3}$$

Precision menjelaskan keakuratan model dalam mengidentifikasi kelas positif dari total prediksi kelas positif yang dilakukan, dihitung sebagai rasio antara prediksi positif yang benar (TP) dengan total prediksi positif (TP + FP). *Precision* berguna dalam kondisi di mana FP lebih dikhawatirkan daripada FN.

$$Presisi = \frac{TP}{TP + FP} \tag{2.4}$$

Recall menjelaskan kemampuan model dalam mengidentifikasi kelas positif dengan benar, dihitung dengan membagi jumlah prediksi yang benar (TP) dengan total kasus positif aktual (TP + FN). *Recall* berguna untuk model yang mengutamakan hasil prediksi positif daripada negatif, seperti model klasifikasi risiko diabetes.

$$Recall = \frac{TP}{TP + FN} \tag{2.5}$$

F1-Score merupakan rata-rata nilai *precision* dan *recall*. Metrik ini sangat berguna dalam konteks *imbalanced data* karena memberikan penilaian yang lebih seimbang terhadap performa model. *F1-Score* yang tinggi mengindikasikan bahwa

model mampu memberikan prediksi yang akurat (*precision* tinggi) sekaligus mendeteksi sebagian besar kasus positif (*recall* tinggi).

$$F1\ Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \tag{2.6}$$

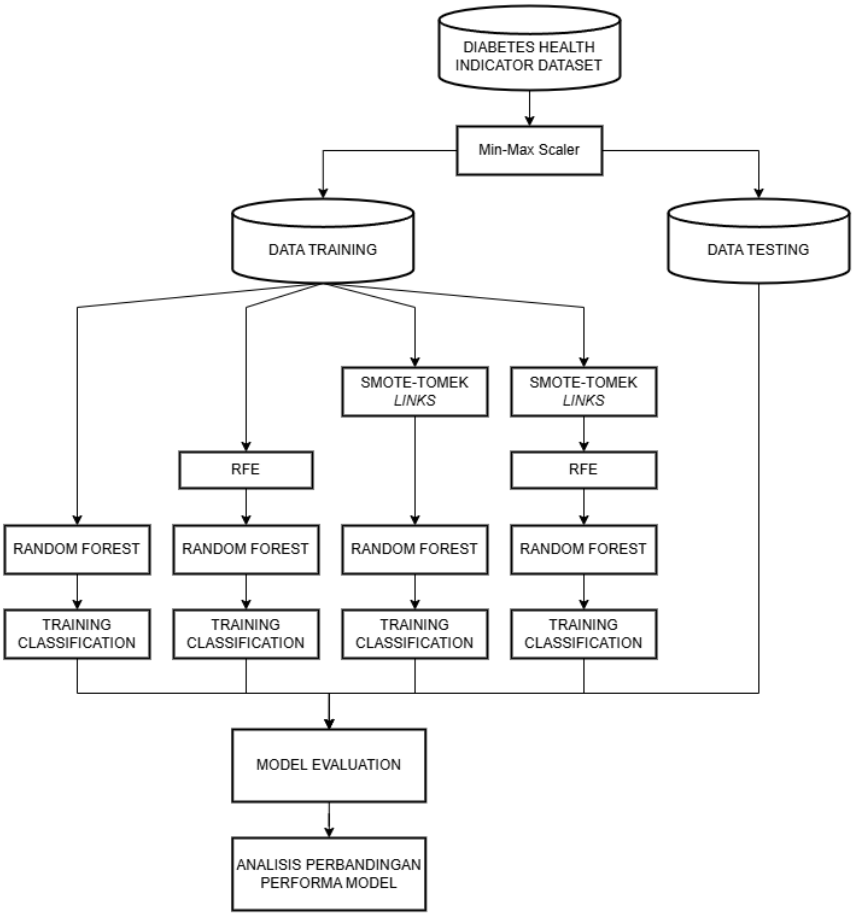
Pemilihan metrik evaluasi yang tepat harus mempertimbangkan konsekuensi dari berbagai jenis kesalahan prediksi dalam konteks spesifik. Dalam kasus prediksi risiko diabetes, *false negative* (gagal mendeteksi pasien berisiko) dianggap memiliki konsekuensi yang lebih serius dibandingkan *false positive*, sehingga *recall* perlu diberi bobot yang lebih tinggi dalam evaluasi model.

BAB III

METODOLOGI PENELITIAN

3.1 Desain Sistem

Desain sistem penelitian ini disusun untuk mengevaluasi pengaruh metode *balancing data* dan seleksi fitur terhadap performa model klasifikasi risiko diabetes menggunakan algoritma *Random Forest*. Sistem ini dibangun dalam beberapa tahap, dengan empat skenario model untuk menentukan kombinasi terbaik yang menghasilkan performa optimal.



Gambar 3. 1 Desain Sistem

3.2 *Diabetes Health Indicator Dataset (DHID)*

Dataset yang digunakan dalam penelitian ini adalah *Diabetes Health Indicator Dataset* (DHID), yang merupakan versi bersih dari data survei *Behavioral Risk Factor Surveillance System* (BRFSS) tahun 2015 yang tersedia secara publik di Kaggle. Dataset ini berisi 253.680 data responden dan dikembangkan oleh *Centers for Disease Control and Prevention* (CDC). Instrumen survei BRFSS telah melalui uji teknis, validasi lapangan, serta uji kognitif untuk memastikan kualitas dan konsistensinya.

Tabel 3. 1 Deskripsi fitur

Fitur	Deskripsi
Diabetes_binary	Kelas target indikator diabetes: 0 untuk tidak diabetes dan 1 untuk pra diabetes atau diabetes
HighBP	Tekanan darah tinggi: 0 = tidak ada tekanan darah tinggi, 1 = memiliki tekanan darah tinggi.
HighChol	Kolesterol tinggi: 0 = tidak ada kolesterol tinggi, 1 = memiliki kolesterol tinggi.
CholCheck	Pemeriksaan kolesterol dalam 5 tahun terakhir: 0 = tidak pernah diperiksa, 1 = pernah diperiksa.
BMI	*Body Mass Index* (Indeks Massa Tubuh).
Smoker	Pernah merokok setidaknya 100 batang sepanjang hidupnya: 0 = tidak, 1 = ya.
Stroke	Pernah didiagnosis stroke: 0 = tidak, 1 = ya.
HeartDiseaseorAttack	Pernah didiagnosis penyakit jantung koroner (CHD) atau serangan jantung (MI): 0 = tidak, 1 = ya.
PhysActivity	Aktivitas fisik dalam 30 hari terakhir (selain pekerjaan): 0 = tidak, 1 = ya.
Fruits	Mengonsumsi buah setidaknya satu kali sehari: 0 = tidak, 1 = ya.
Veggies	Mengonsumsi sayuran setidaknya satu kali sehari: 0 = tidak, 1 = ya.
HvyAlcoholConsump	Konsumsi alkohol berat (untuk pria dewasa ≥14 gelas per minggu dan wanita dewasa ≥7 gelas per minggu): 0 = tidak, 1 = ya.
AnyHealthcare	Memiliki jenis jaminan kesehatan seperti asuransi atau HMO: 0 = tidak, 1 = ya.
NoDocbcCost	Tidak bisa pergi ke dokter dalam 12 bulan terakhir karena biaya: 0 = tidak, 1 = ya.

Fitur	Deskripsi
GenHlth	Kesehatan umum menurut diri sendiri: Skala 1-5, 1 = sangat baik, 2 = baik sekali, 3 = baik, 4 = cukup, 5 = buruk.
MentHlth	Jumlah hari mengalami kesehatan mental yang buruk dalam 30 hari terakhir (skala 1-30 hari).
PhysHlth	Jumlah hari mengalami penyakit fisik atau cedera dalam 30 hari terakhir (skala 1-30 hari).
DiffWalk	Kesulitan serius dalam berjalan atau menaiki tangga: 0 = tidak, 1 = ya.
Sex	Jenis kelamin: 0 = perempuan, 1 = laki-laki.
Age	Kategori umur: 1 = umur 18-24 tahun, 2 = umur 25-29 tahun, 3 = umur 30-34 tahun, ..., 13 = umur 80 tahun ke atas.
Education	Tingkat pendidikan: 1 = tidak pernah bersekolah atau hanya TK, 2 = sekolah dasar, 3 = sekolah menengah pertama, 4 = sekolah menengah atas, 5 = sebagian perguruan tinggi, 6 = lulus perguruan tinggi.
Income	Penghasilan. Berskala 1-8: 1 = < 10k, 2 = 10k-15k, 3 = 15k-20k, ..., 8 > 75k.

Tabel 3.1 merupakan penjelasan fitur-fitur yang ada pada *dataset*. Setiap fitur dalam *dataset* ini dikategorikan dan disusun berdasarkan pendekatan survei kesehatan masyarakat dan praktik epidemiologi global. Struktur data mengikuti sistem pengukuran yang dirancang untuk merepresentasikan indikator risiko secara terstandarisasi.

Dasar pengkategorisasian fitur-fitur tersebut memiliki landasan ilmiah yang kuat. Sebagai contoh, variabel umur dalam BRFSS hanya mencakup individu berusia 18 tahun ke atas karena survei ini memang dirancang untuk populasi dewasa, sesuai pedoman CDC dan WHO. Hal ini sejalan dengan fakta bahwa diabetes tipe 2 umumnya menyerang usia dewasa dan berkaitan erat dengan faktor risiko seperti hipertensi, obesitas, dan pola hidup tidak sehat (American Diabetes Association, 2020).

Variabel *Education* dan *Income* dikategorikan berdasarkan tingkat pendidikan terakhir dan rentang penghasilan tahunan. Pendekatan ini sesuai dengan prinsip *social determinants of health*, di mana status pendidikan dan ekonomi memengaruhi akses terhadap layanan kesehatan, pengetahuan tentang gizi, serta peluang untuk hidup sehat.

Validitas struktur dan isi data BRFSS telah dibuktikan oleh banyak penelitian. Studi oleh (Pierannunzi, Hu, & Balluz, 2013) menunjukkan bahwa hasil BRFSS sebanding dengan survei nasional lain seperti NHIS dan NHANES. Penelitian lain seperti (Wang dkk., 2021) juga berhasil menggunakan data BRFSS untuk membangun model prediksi diabetes dengan hasil yang signifikan secara statistik.

Berdasarkan penjelasan di atas, DHID dapat dianggap sebagai sumber data yang valid, terpercaya, dan relevan untuk keperluan pengembangan model klasifikasi risiko diabetes dalam penelitian ini.

3.3 Normalisasi Data

Dataset dinormalisasi terlebih dahulu untuk menyeragamkan nilai dari berbagai fitur ke dalam rentang nilai yang sama. Metode normalisasi yang digunakan merupakan *Min-Max Scaler.* Metode ini merupakan teknik penskalaan yang mengubah nilai fitur X ke dalam rentang [0,1] menggunakan rumus transformasi sebagai berikut:

$$X_{normalize} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

Di mana X_{min} adalah nilai minimum dari fitur yang diamati, dan Xmax adalah nilai maksimum dari fitur yang diamati. Keunggulan Min-Max *Scaler* terletak pada kemampuannya untuk memetakan data secara linier ke dalam rentang yang telah ditentukan.

Melalui penerapan Min-Max *Scaler* ini, seluruh fitur numerik dalam *dataset* telah disesuaikan ke dalam rentang nilai [0,1], sehingga meminimalkan bias yang disebabkan oleh perbedaan skala dan mempersiapkan data untuk pemodelan *machine learning* yang optimal.

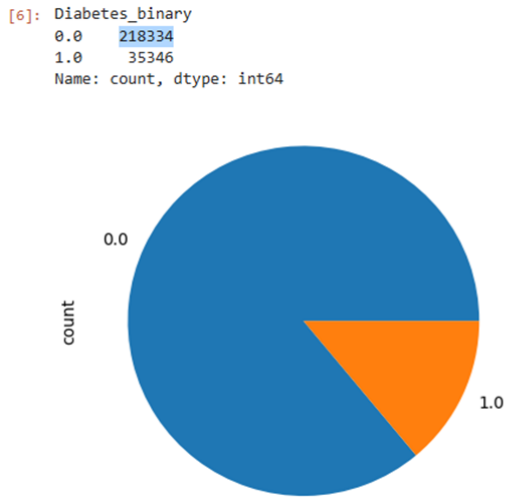
3.4 *Split* Data

Dataset pada penelitian ini dipisah menjadi 2 bagian yaitu, data latih dan data uji dengan rasio 70:30. Pemilihan rasio 70:30 didasarkan pada pertimbangan keseimbangan antara jumlah data latih dan data uji. Bagian yang lebih besar dari *dataset* digunakan untuk melatih model agar dapat mengenali pola dalam data dengan lebih baik. Semakin banyak data latih, semakin baik kemungkinan model dapat mempelajari hubungan kompleks di dalam dataset. Sebagian data dipisahkan sebagai data uji untuk mengukur kemampuan model dalam membuat prediksi pada data yang belum pernah dilihat sebelumnya. Rasio 30% dianggap cukup untuk memberikan evaluasi performa yang akurat tanpa mengurangi data latih secara berlebihan.

Penggunaan rasio 70:30 ini juga sesuai dengan praktik umum dalam *machine learning,* terutama pada *dataset* berukuran besar seperti *Diabetes Health Indicator Dataset* yang digunakan dalam penelitian ini. Dengan jumlah data latih

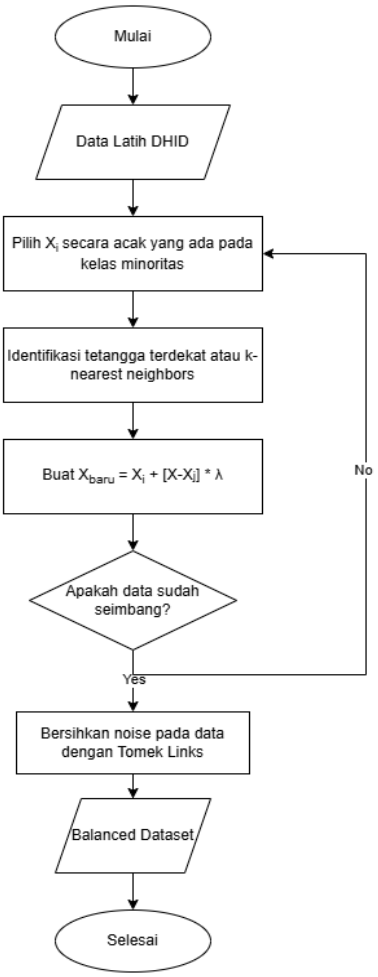
yang mencukupi, model dapat belajar dengan baik, sementara data uji yang cukup banyak memastikan evaluasi model tidak bias terhadap data latih.

3.5 SMOTE-Tomek *Links*



Gambar 3. 2 Diagram distribusi kelas

SMOTE-Tomek *Links* digunakan untuk menangani ketidakseimbangan data. *Dataset* yang digunakan pada penelitian ini memiliki ketidakseimbangan data, seperti yang terlihat pada gambar 3.2 jumlah data dengan kelas 0 (tidak diabetes) sebanyak 218.334, sedangkan data dengan kelas 1 (diabetes) hanya 35.346. SMOTE-Tomek *Links* pada penelitian ini digunakan setelah proses *split data* dan hanya digunakan pada data latih.



Gambar 3. 3 *Flowchart* SMOTE-TOMEK *Links*

Gambar 3.3 merupakan *flowchart* yang berisi langkah-langkah SMOTE-Tomek *Links* dalam menyeimbangkan data latih DHID pada penelitian ini. Adapun penjelasan langkah-langkahnya sebagai berikut:

1. Tahap pertama yang dilakukan adalah membagi data yang telah dinormalisaikan menjadi 2 bagian yaitu data latih dan data uji. Penyeimbangan data menggunakan SMOTE-Tomek *Links* ini hanya diterapkan pada data latih.
2. Pilih X_i atau data acak yang ada pada kelas minoritas.

3. Identifikasi tetangga terdekat dari data X_i
4. SMOTE akan membuat data sintetis baru dengan cara melakukan interpolasi antara data asli dengan tetangga terplihnya.

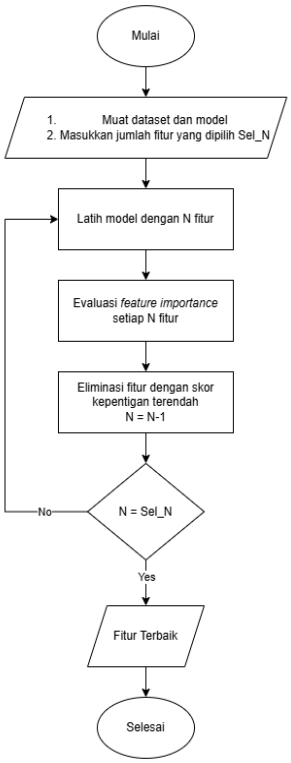
$$X_{baru} = X_i + [X - X_j] \times \lambda \quad (3.2)$$

Keterangan :
X_{baru} : data sintetis baru yang dihasilkan
X_i : sampel kelas minoritas yang dipilih
X_j : salah satu tetangga terdekat dari X_i (kelas minoritas)
λ : bilangan acak rentang 0 dan 1

5. Proses ini diulang sampai jumlah data kelas minoritas seimbang dengan jumlah data kelas mayoritas
6. Setelah itu, Tomek *Links* akan membersihkan *noise* pada data dengan cara menghapus *sample* data kelas mayoritas yang berdekatan dengan *sample* data kelas minoritas.
7. Dataset berhasil diseimbangkan, algoritma selesai.

3.6 *Recursive Feature Elemination* (RFE)

Dalam penelitian ini, *Recursive Feature Elemination* (RFE) digunakan sebagai metode seleksi fitur untuk mengidentifikasi fitur-fitur yang paling berpengaruh dalam klasifikasi risiko diabetes. RFE menggunakan prinsip *backward elimination*, yaitu dimulai dengan semua fitur yang tersedia, kemudian secara iteratif menghapus fitur-fitur yang dianggap paling tidak penting hingga mencapai jumlah fitur yang diinginkan.



Gambar 3. 4 *Flowchart* RFE

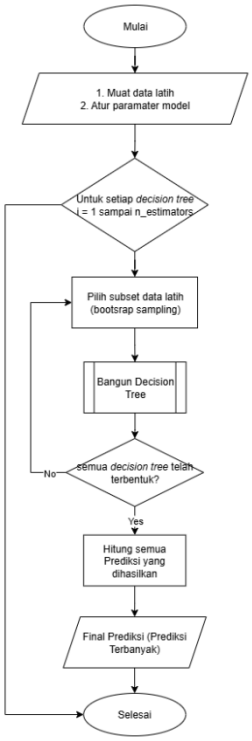
Flowchart di gambar 3.4 menjelaskan alur proses RFE, dari tahap pelatihan awal dengan semua fitur hingga pemilihan *subset* fitur terbaik. Adapun penjelasan langkah-langkahnya sebagai berikut:

1. Tahap pertama yang dilakukan adalah menyiapkan model *Random Forest* sebagai estimator untuk menentukan *feature importance*, menyiapkan *dataset* yang telah diseimbangkan menggunakan SMOTE-Tomek *Links*, dan menentukan jumlah fitur yang akan diseleksi (Sel_N).
2. Selanjutnya RFE melatih model *Random Forest* menggunakan semua fitur (N) pada *dataset*.

3. Setelah pelatihan, RFE mengevaluasi *feature importance* untuk setiap fitur dalam model.
4. Fitur dengan skor *importance* terendah dihapus untuk mengurangi gangguan dari fitur yang tidak relevan (N = N-1).
5. Proses ini diulang hingga hanya tersisa jumlah fitur optimal yang diinginkan (N = sel_F).
6. Setelah N = sel_F algoritma akan menghasilkan data dengan *subset* fitur terbaik.

Subset fitur ini kemudian digunakan untuk melatih model *Random Forest* dalam langkah berikutnya, sehingga diharapkan dapat meningkatkan akurasi dalam klasifikasi risiko diabetes.

3.7 *Random Forest*



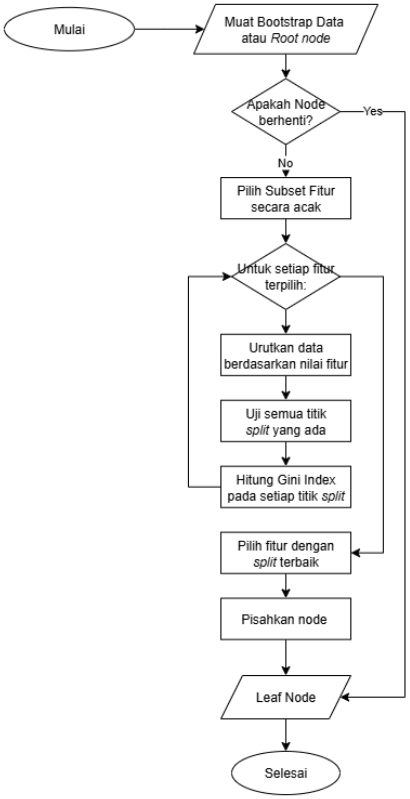
Gambar 3. 5 *Flowchart Random Forest*

Random Forest bekerja dengan membangun sejumlah pohon keputusan (*decision tree*) dan menggabungkan hasil prediksi dari masing-masing pohon menggunakan teknik voting mayoritas untuk tugas klasifikasi.

Flowchart pada gambar 3.5 menjelaskan bagaimana algoritma random forest bekerja. Adapun penjelasan langkah-langkahnya sebagai berikut:

1. Langkah pertama adalah memuat data latih dan menentukan paramater algoritma, seperti: n_estimators untuk jumlah *decision tree* yang akan dibangun dan max_depth untuk jumlah kedalaman maksimum setiap *decision tree*.
2. Untuk setiap *decision tree* i = 1 sampai n_estimators adalah *loop* utama. Algoritma akan mengulang langkah-langkah berikut sebanyak n_estimators kali untuk membangun setiap *decision tree* yang akan menjadi bagian dari hutan.
3. Kemudian dilakukan sampling acak dengan pengembalian (*Bootstrap sampling)* dari data latih. Hal ini memastikan bahwa setiap pohon dilatih pada variasi data yang sedikit berbeda, yang penting untuk keragaman hutan.
4. Bangun *Decision Tree* berdasarkan data hasil *bootstrap* dengan langkah-langkah yang dijelaskan pada gambar 3.6
5. Cek apakah semua *Decision Tree* telah terbentuk? jika iya maka ulangi dari langkah ke-3.

6. Setelah semua *Decision Tree* terbentuk, prediksi terhadap data uji dilakukan dengan mengambil suara dari masing-masing pohon, lalu hasil akhir ditentukan berdasarkan kelas yang paling banyak dipilih (mayoritas).
7. Algoritma selesai.



Gambar 3. 6 *Flowchart Decision Tree*

Gambar 3.6 menjelaskan bagaimana *Decision Tree* dibangun. Sebuah Pohon dibangun secara rekursif dengan pemilihan *subset* fitur dan pembagian *node* berdasarkan kriteria *Gini Impurity*. Berikut penjelasan alurnya:

1. Proses dimulai dari *root node* (*node* akar) yang berisi seluruh data hasil *bootstrap*.

2. Pada setiap *node*, algoritma memilih sejumlah fitur secara acak dari keseluruhan fitur yang tersedia. Tujuannya untuk meningkatkan keragaman pohon.

$$m = \sqrt{p} \tag{3.3}$$

Keterangan :
p : jumlah total fitur

3. Urutkan data dan uji semua titik *split* pada setiap fitur yang terpilih
4. Hitung *gini* indeks pada setiap titik *split.* Setiap titik *split* yang diuji, *Gini Impurity* dari *node* anak yang akan dihasilkan dihitung. Tujuannya adalah untuk menemukan *split* yang paling "murni" atau menghasilkan penurunan *Gini Impurity* terbesar (*Gini Gain*). Untuk menghitung *Gini Impurity* dapat menggunakan persamaan 2.1.
5. Hitung *Gini Split.* Setiap titik split menghasilkan dua node anak. Gini Split adalah rata-rata tertimbang dari Gini Impurity di kedua node anak tersebut. Untuk menghitung *Gini Split* dapat menggunakan persamaan 2.2. *Gini Gain* kemudian dapat dihitung dengan persamaan:

$$Gini\_Gain = Gini_{parent} - Gini\_split \tag{3.4}$$

6. Pilih fitur dengan *split* terbaik. Dari semua kombinasi fitur acak dan titik *split*, algoritma memilih fitur dan titik *split* yang menghasilkan *Gini Gain* tertinggi. *Split* ini akan digunakan untuk membagi *node*.
7. Pada tahap pisahkan *node*, *node* akan dibagi menjadi dua cabang dari *node* berdasarkan fitur dan nilai *split* yang terpilih.

8. Jika tidak bisa atau tidak perlu dipecah lagi, *node* menjadi *leaf* dan tentukan kelas prediksi (mayoritas kelas di *node*).
9. *Decision Tree* selesai dibangun.

3.8 Implementasi Model

Pada tahap ini akan dilakukan pelatihan dan pengujian model menggunakan algoritma *Random Forest* untuk klasifikasi risiko diabetes. Adapun alur implementasinya sebagai berikut:

1. Data *input* yang digunakan pada tahap ini telah melalui proses *pre-processing* SMOTE-Tomek *Links* dengan subset fitur yang dihasilkan RFE.
2. *Tuning hyperparamater* menggunakan *GridSeacrhCV* juga dilakukan agar mendapatkan kombinasi paramater yang optimal. Pada penelitian ini paramater yang digunakan seperti pada tabel 3.3 berikut:

Tabel 3. 2 Tuning paramater

Paramater	Keterangan	Value
n_estimators	Jumlah pohon dalam hutan.	100, 300
Max_depth	Jumlah kedalaman pohon.	5, 10, none

3. Setelah itu model *Random Forest* akan dilatih menggunakan data dan kombinasi *hyperparamater* tersebut.
4. Setelah pelatihan, model dievaluasi menggunakan beberapa metrik evaluasi seperti nilai akurasi, *precision*, *recall,* dan *f1-score.*

3.9 Evaluasi Model

Model yang telah dilatih akan dievaluasi untuk mengukur keakuratannya dalam mengklasifikasikan risiko diabetes. Evaluasi dilakukan menggunakan

metode *confusion matrix* dengan metrik yang banyak digunakan dalam masalah klasifikasi medis, yaitu akurasi, *precision*, *recall*, dan *f1-score.*

3.10 Skenario Uji Coba

Peneliti melakukan skenario uji coba untuk mengevaluasi performa model klasifikasi risiko diabetes dengan menerapkan berbagai pendekatan dalam penanganan *dataset* dan seleksi fitur. Hal ini dilakukan untuk mengetahui kombinasi mana yang menghasilkan performa terbaik berdasarkan metrik evaluasi. Adapun skenario uji coba dapat dilihat pada tabel 3.3.

Tabel 3. 3 Skenario model

Skenario	**Penanganan *Dataset***	**Seleksi Fitur**
Skenario 1	-	-
Skenario 2	-	RFE
Skenario 3	SMOTE-Tomek *Links*	-
Skenario 4	SMOTE-Tomek *Links*	RFE

Pada skenario 1 model dilatih menggunakan *dataset* asli tanpa penyeimbangan kelas dan seleksi fitur. Dalam skenario 2, model dilatih menggunakan *dataset* asli dan seleksi fitur menggunakan RFE. Skenario 3 dan 4 model dilatih dengan *dataset* yang telah diseimbangkan dengan SMOTE-Tomek *Links*, akan tetapi pada skenario 4 dilakukan juga seleksi fitur menggunakan RFE.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil Pengujian

Pada bab ini akan dijelaskan hasil pengujian dari semua tahapan mulai dari *split data,* SMOTE-Tomek *Links*, RFE, implementasi *Random Forest,* hingga pengujian model skenario.

4.1.1 Normalisasi Data

Implementasi normalisasi pada penelitian ini dilakukan menggunakan metode Min-Max Scaler, yaitu teknik penskalaan yang mengubah nilai fitur ke dalam rentang [0,1]. Pilihan Min-Max Scaler didasarkan pada kesederhanaan dan efektivitasnya dalam memetakan data secara linier, yang sesuai untuk memastikan bahwa semua fitur berkontribusi secara proporsional. Berikut adalah contoh hasil implementasi normalisasi pada *dataset*:

Tabel 4. 1 Data sebelum normalisasi

HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsumn	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Target
1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0	0.0
0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0	0.0
1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0	0.0
1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0	0.0
1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0	0.0

Tabel 4. 2 Data setelah normalisasi

HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsumn	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income	Target
0.0	1.0	1.0	0.32	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.25	0.6	0.5	1.0	0.0	0.66	0.6	0.28	0.0
0.0	0.0	1.0	0.15	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.50	1.0	0.0	0.0
1.0	1.0	1.0	0.18	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.5	1.0	0.1	1.0	0.0	0.66	0.6	1.0	0.0
0.0	1.0	1.0	0.17	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.5	0.0	0.0	0.0	0.0	0.83	0.4	0.71	0.0
0.0	0.0	1.0	0.13	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	0.5	0.1	0.0	0.0	0.0	0.83	0.8	0.42	0.0

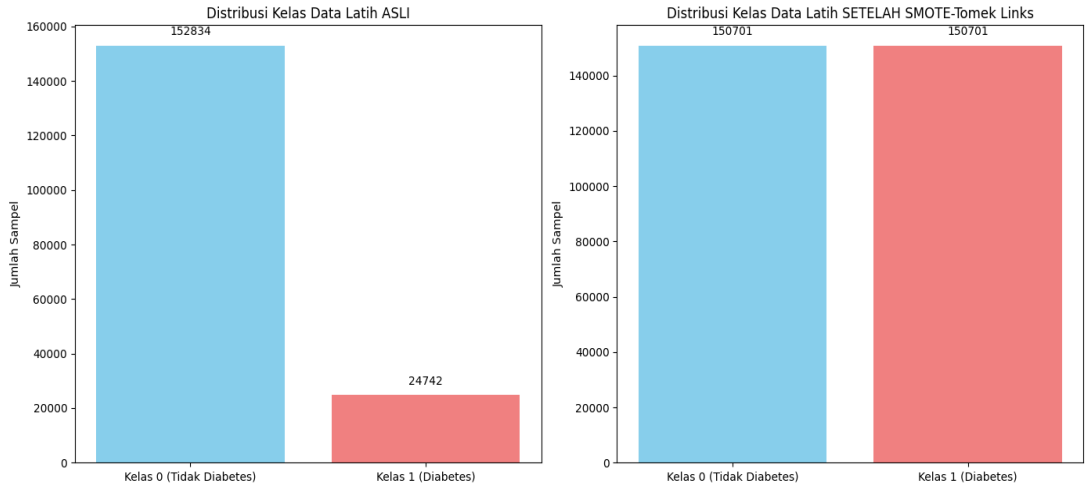
4.1.2 Hasil Split Data

Pada proses *split data*, dataset dibagi menjadi dua bagian yaitu, data latih dan data uji dengan rasio 70:30. Pemilihan rasio 70:30 didasarkan pada pertimbangan keseimbangan antara jumlah data latih dan data uji. Berikut tabel yang menampilkan hasil *split* data dengan rasio 70:30.

Tabel 4. 3 Hasil *Split Data*

Jumlah Data Latih	Jumlah Data Uji
177.576	76.104

4.1.3 SMOTE-Tomek *Links*



Gambar 4. 1 Hasil SMOTE-Tomek Links

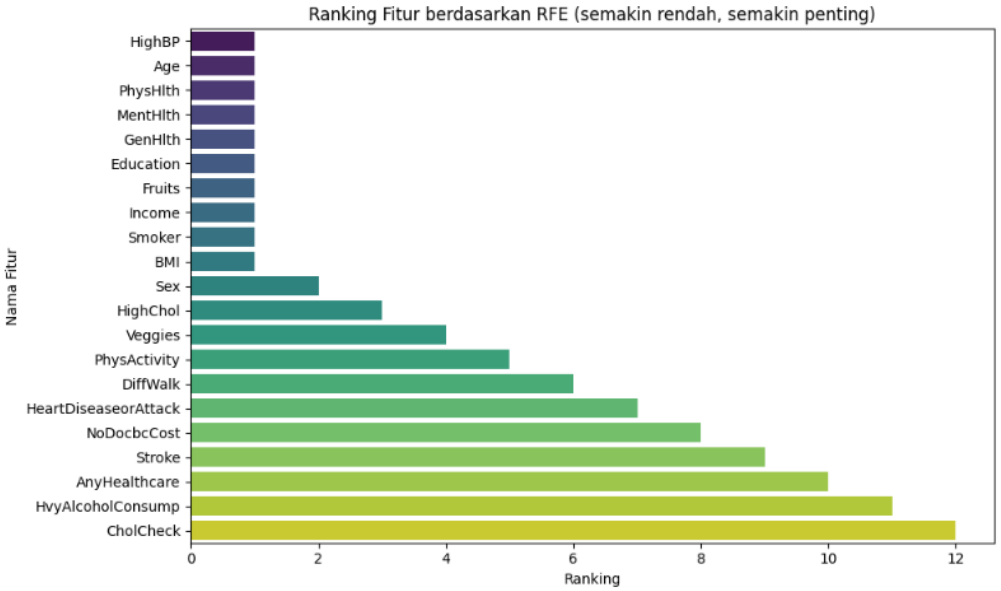
Gambar 4.1 menunjukkan distribusi kelas sebelum dan sesudah dilakukan SMOTE-Tomek *Links.* Distribusi sebelum *resampling* menunjukkan bahwa kelas 0 (tidak diabetes) berjumlah 152.834 data, sedangkan kelas 1 (diabetes) hanya berjumlah 24.742 data. Ketidakseimbangan ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas. Sedangkan distribusi setelah dilakukan SMOTE-Tomek *Links* menunjukkan bahwa kedua kelas memiliki jumlah data yang

seimbang, masing-masing 150.701 data untuk kelas 0 (tidak diabetes) dan 150.701 data untuk kelas 1 (diabetes).

4.1.4 *Recursive Feature Elimination (RFE)*

Untuk meningkatkan performa model dan mengurangi kompleksitas komputasi, dilakukan proses seleksi fitur menggunakan metode *Recursive Feature Elimination* (RFE). RFE bekerja dengan cara membangun model dan mengeliminasi fitur yang paling tidak penting secara berulang-ulang hingga jumlah fitur yang ditentukan tercapai. Dalam penelitian ini, 10 fitur terbaik akan dipilih dan algoritma *Random Forest* sebagai estimator dasar untuk proses seleksi fitur.

RFE diterapkan pada dataset hasil *resampling* dengan SMOTE-Tomek *Links*. Proses ini dilakukan untuk memilih 10 fitur terbaik dari keseluruhan fitur yang tersedia.



Gambar 4. 2 Rangking Seleksi Fitur RFE

Gambar 4.2 menunjukkan rangking setiap fitur yang telah dihitung oleh RFE, 10 fitur dengan rangking 1 merupakan fitur terbaik yang dipilih RFE dari keseluruhan fitur. Hasil fitur yang terpilih oleh RFE adalah sebagai berikut: *HighBP, BMI, Smoker, Fruits, GenHlth, MentHlth, PhysHlth,Age, Education, Income.*

4.1.5 Hasil Uji Coba Model Skenario

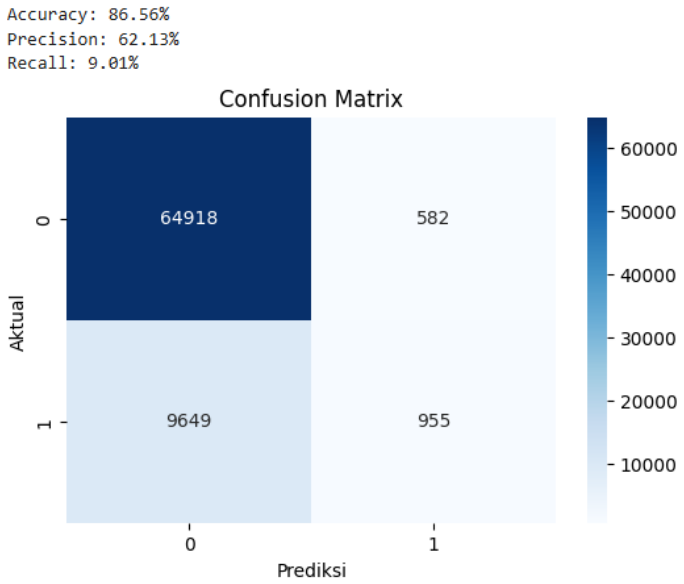
Untuk mengevaluasi pengaruh dari penanganan data yang tidak seimbang dan seleksi fitur terhadap performa model, dilakukan pengujian pada empat skenario berbeda seperti yang telah dijelaskan pada Tabel 3.3. Pada setiap skenario, model *Random Forest* tidak langsung digunakan dengan parameter default.

Untuk memperoleh performa terbaik dari model *Random Forest* pada setiap skenario, dilakukan proses *hyperparameter tuning* menggunakan GridSearchCV. GridSearchCV digunakan untuk mencari kombinasi parameter terbaik dengan teknik *cross-validation* sebanyak 5 lipatan (*5-fold cross-validation*). Adapun paramater yang disesuaikan seperti pada Tabel 3.2. Melalui proses ini, diperoleh kombinasi parameter optimal untuk masing-masing skenario yang kemudian digunakan untuk pelatihan model dan evaluasi.

Setelah model dilatih menggunakan parameter terbaik, dilakukan evaluasi performa menggunakan *confusion matrix*. *Confusion matrix* memberikan gambaran seberapa baik model mampu mengklasifikasikan dua kelas pada dataset, yaitu kelas 0 (tidak berisiko diabetes) dan kelas 1 (berisiko diabetes). Dari *confusion matrix* diperoleh metrik evaluasi seperti: akurasi, presisi, *recall*, dan f1-*score*. Berikut adalah hasil uji coba masing-masing skenario:

4.1.5.1 Skenario Model Tanpa Penanganan Data dan Seleksi Fitur

Pada skenario ini model dibangun langsung menggunakan dataset asli tanpa melalui proses penyeimbangan maupun seleksi fitur. Pada saat *hyperparamater tuning* menggunakan GridSearchCV model ini mendapatkan hasil terbaik pada kombinasi paramater *max_depth* = 10 dan *n_estimators* = 300, dengan *confusion matrix* seperti pada gambar berikut:



Gambar 4. 3 *Confusion Matrix* Skenario Model 1

Pada gambar 4.3 model menghasilkan *True Positive* (TP) = 955 data pasien berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi benar). *False Positive* (FP) = 582 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi salah), *True Negative* (TN) = 64.918 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi benar), *False Negative* (FN) = 9.649 data pasien berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi salah).

Dari *confusion matrix* di atas dapat diketahui skor akurasi, presisi, recall, dan f1-score yang didapat model tanpa penanganan data dan seleksi fitur.

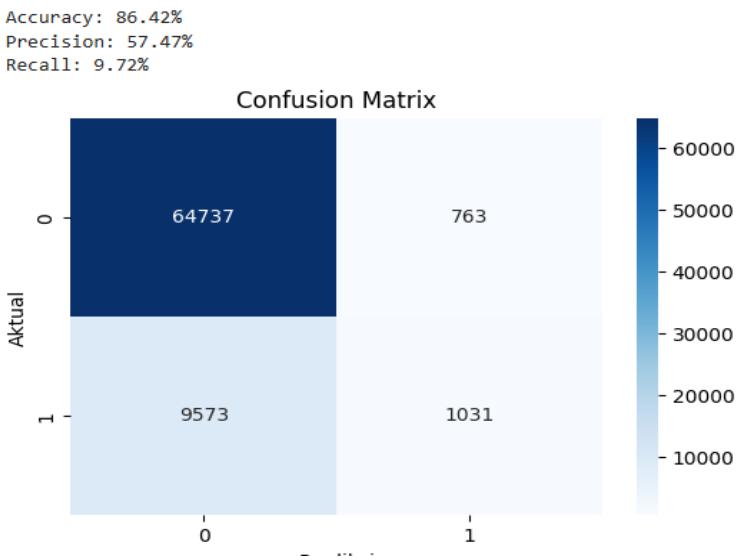
Tabel 4. 4 matrik evaluasi skenario model 1

Akurasi	Presisi	Recall	F1-Score
86,56%	62,13%	9,01%	15,73%

4.1.5.2 Skenario Model dengan RFE

Pada skenario model kedua dibangun dengan melalui proses seleksi fitur RFE untuk memilih 10 fitur terbaik, fitur-fitur tersebut adalah: *HighBP, BMI, Smoker, Fruits, GenHlth, MentHlth, PhysHlth, Age, Education, Income*, Urutan *rangking* sebagaimana terlihat pada gambar 4.2

Pada saat *hyperparamater tuning* menggunakan GridSearchCV model ini mendapatkan hasil terbaik pada kombinasi paramater *max_depth* = 10 dan *n_estimators* = 300, dengan *confusion matrix* seperti pada gambar berikut:



Gambar 4. 4 *Confusion Matrix* Skenario Model 2

Pada gambar 4.4 model menghasilkan *True Positive* (TP) = 1.031 data pasien berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi

benar). *False Positive* (FP) = 763 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi salah), *True Negative* (TN) = 64.737 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi benar), *False Negative* (FN) = 9.573 data pasien berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi salah).

Dari *confusion matrix* di atas dapat diketahui skor akurasi, presisi, *recall*, dan f1-*score* yang didapat skenario model dengan seleksi 10 fitur menggunakan metode RFE, berikut adalah hasilnya:

Tabel 4. 5 Matrik evaluasi skenario model 2

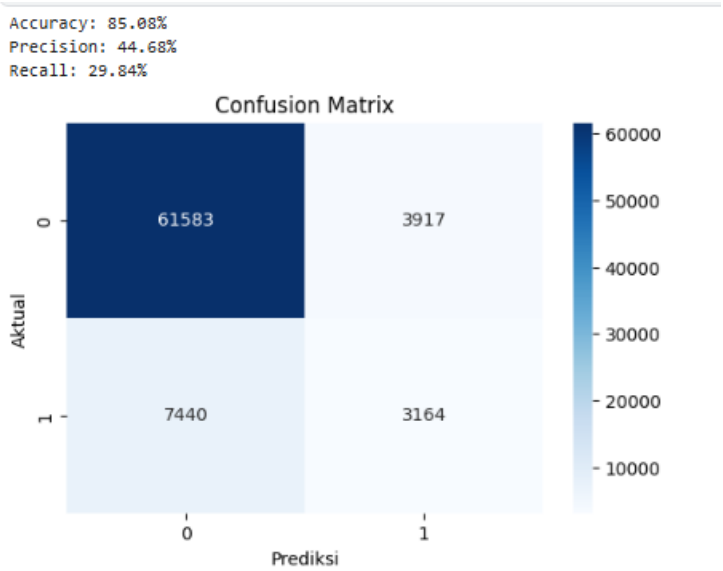
Akurasi	Presisi	Recall	F1-Score
86,42%	57,47%	9,72%	16,63%

x

4.1.5.3 Skenario Model dengan SMOTE-Tomek Links

Pada skenario ketiga, *dataset* terlebih dahulu diproses menggunakan SMOTE-Tomek *Links* untuk menyeimbangkan distribusi kelas, adapun hasil penyeimbangan data pada model ini bisa dilihat pada gambar 4.1.

Setelah itu, dilakukan proses *hyperparameter tuning* menggunakan GridSearchCV untuk mendapatkan kombinasi parameter terbaik. Model ini mendapatkan hasil terbaik pada kombinasi paramater *max_depth* = *None* dan *n_estimators* = 300, dengan *confusion matrix* seperti pada gambar berikut:



Gambar 4. 5 *Confussion Matrix* Skenario Model 3

Pada gambar 4.5 model menghasilkan True Positive (TP) = 3.164 data pasien berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi benar). False Positive (FP) = 3.917 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi salah), True Negative (TN) = 61.583 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi benar), False Negative (FN) = 7.440 data pasien berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi salah).

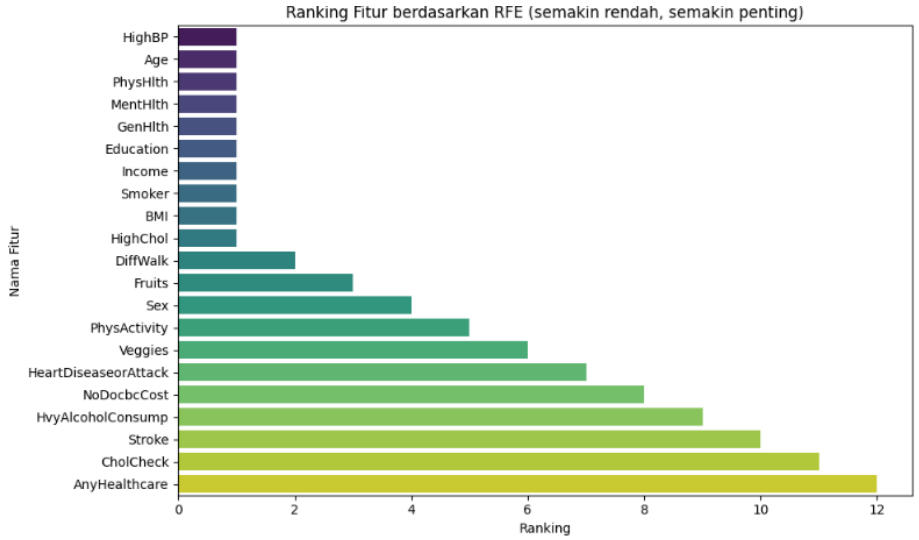
Berdasarkan Gambar 4.5 dapat diketahui skor akurasi, presisi, *recall*, dan f1-*score* yang didapat skenario model 3 dengan penanganan data SMOTE-Tomek *Links*, berikut adalah hasilnya:

Tabel 4. 6 Matrik evaluasi skenario model 3

Akurasi	**Presisi**	**Recall**	**F1-Score**
85,08%	44,68%	29,84%	35,78%

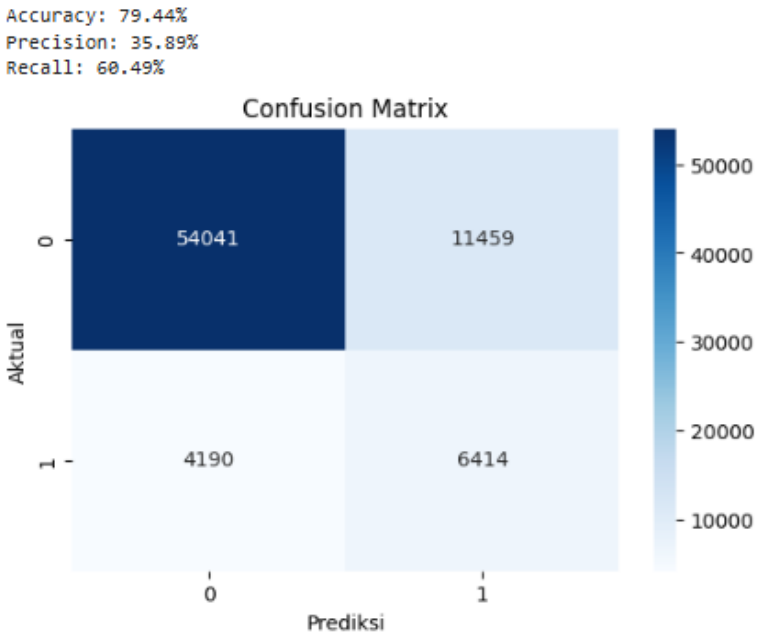
4.1.5.4 Skenario Model dengan SMOTE-Tomek *Links* dan RFE

Pada skenario keempat, *dataset* terlebih dahulu diproses menggunakan SMOTE-Tomek *Links* untuk menyeimbangkan distribusi antar kelas, serta menggunakan metode RFE untuk memilih 10 fitur terbaik. Gambar 4.1 terlihat distribusi kelas 1 dan 0 seimbang setelah dilakukan proses SMOTE-Tomek *Links*, sedangkan pada Gambar 4.6 menunjukkan rangking fitur terpenting hasil dari proses RFE.



Gambar 4. 6 Hasil seleksi fitur RFE pada model 4

Setelah itu, dilakukan proses *hyperparameter tuning* menggunakan GridSearchCV untuk mendapatkan kombinasi parameter terbaik. Parameter terbaik yang diperoleh adalah *max_depth* = 10 dan *n_estimators* = 300, dengan *confusion matrix* seperti pada gambar berikut:



Gambar 4. 7 *Confusion Matrix* Skenario Model 4

Pada gambar 4.7 model menghasilkan *True Positive* (TP) = 6.414 data pasien berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi benar). *False Positive* (FP) = 11.459 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien berisiko diabetes (prediksi salah), *True Negative* (TN) = 54.041 data pasien tidak berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi benar), False Negative (FN) = 4.190 data pasien berisiko diabetes yang diprediksi sebagai pasien tidak berisiko diabetes (prediksi salah).

Berdasarkan Gambar 4.10 dapat diketahui skor akurasi, presisi, *recall*, dan f1-score yang didapat skenario model 4 dengan penanganan data SMOTE-Tomek *Links* dan seleksi fitur RFE, berikut adalah hasilnya:

Tabel 4. 7 Matrik evaluasi skenario model 4

Akurasi	**Presisi**	**Recall**	**F1-Score**
79,44%	35,89%	60,49%	45,05%

4.2 Pembahasan Analisis Skenario Model

Pada sub bab ini akan dibahas mengenai hasil uji coba dari implementasi *Random Forest* yang dikombinasikan dengan metode penyeimbangan data SMOTE-Tomek *Links* serta seleksi fitur menggunakan *Recursive Feature Elimination* (RFE). Empat skenario pengujian dibandingkan untuk mengevaluasi dampak penyeimbangan data dan seleksi fitur terhadap performa model.

4.2.1 Analisis Skenario Model 1

Tabel 4. 8 Analisa skenario model 1

Analisa Skenario Model 1		
Best parameter	N_estimators	300
	Max_depth	10
Confusion Matrix	*True Positive*	955
	False Positive	582
	True Negative	64.918
	False Negative	9.649
Matrik Evaluasi	Akurasi	86,56%
	Presisi	62,13%
	Recall	9,01%
	F1-*Score*	16,63%

Pada skenario model 1, model dibangun menggunakan data asli tanpa penanganan ketidakseimbangan data maupun seleksi fitur. Model menghasilkan akurasi tinggi (86,56%) namun *recall* rendah (9,01%). Hal ini menunjukkan bahwa model cenderung bias terhadap kelas mayoritas (non-diabetes), sehingga kemampuan mendeteksi pasien berisiko diabetes sangat buruk, meskipun akurasi terlihat baik secara keseluruhan.

Ketidakseimbangan data merupakan masalah umum dalam klasifikasi medis, di mana jumlah kasus negatif (non-diabetes) jauh lebih banyak dibandingkan kasus positif (diabetes). Dalam kondisi seperti ini, model cenderung memprediksi kelas mayoritas, sehingga metrik akurasi bisa tampak tinggi, padahal sebenarnya

kemampuan mendeteksi kasus minoritas sangat buruk (Septiana Rizky, Haiban Hirzi, & Hidayaturrohman, 2022).

Selain itu, algoritma *machine learning* konvensional seperti *Random Forest* tanpa penanganan ketidakseimbangan data akan mengalami penurunan sensitivitas terhadap kelas minoritas. Hal ini terlihat pada hasil *recall* yang sangat rendah, menunjukkan bahwa pasien dengan risiko diabetes gagal terdeteksi oleh model.

4.2.2 Analisis Skenario Model 2

Tabel 4. 9 Analisa skenario model 2

Analisa Skenario Model 2		
Best parameter	N_estimators	100
	Max_depth	10
Confusion Matrix	*True Positive*	1.031
	False Positive	717
	True Negative	64.737
	False Negative	9.573
Matrik Evaluasi	Akurasi	86,42%
	Presisi	57,47%
	Recall	9,72%
	F1-*Score*	16,63%

Skenario model 2 dibangun dengan menggunakan RFE untuk memilih 10 fitur terbaik. Model mendapatkan nilai akurasi 86,42%, presisi 57,47%, *recall* 9,72%, dan f1-*score* 16,63% Jika dibandingkan dengan model pertama, terjadi sedikit peningkatan pada nilai *recall* (dari 9,01% menjadi 9,72%), akan tetapi terjadi sedikit penurunan pada nilai akurasi (dari 86,56% menjadi 86,42%).

Peningkatan performa yang relatif kecil ini menunjukkan bahwa walaupun RFE berhasil menyaring fitur yang paling relevan, hal tersebut belum cukup untuk memperbaiki kemampuan model dalam mendeteksi kasus kelas minoritas (diabetes). Hal ini mengindikasikan bahwa permasalahan utama model terletak

pada ketidakseimbangan distribusi kelas, bukan hanya pada kualitas atau jumlah fitur.

Secara umum, seleksi fitur melalui RFE dapat membantu mengurangi kompleksitas model dan menghindari *overfitting* dengan mengeliminasi fitur yang tidak memberikan kontribusi signifikan terhadap proses klasifikasi. Dalam skenario ini, meskipun jumlah fitur telah dikurangi menjadi 10 fitur yang paling relevan, kinerja *recall* tetap sangat rendah. Hal ini terjadi karena model tetap dilatih pada data yang tidak seimbang, sehingga cenderung mempelajari pola dari kelas mayoritas (non-diabetes) secara lebih dominan.

4.2.3 Analisis Skenario Model 3

Tabel 4. 10 Analisa skenario model 3

Analisa Skenario Model 3		
Best parameter	N_estimators	300
	Max_depth	None
Confusion Matrix	*True Positive*	3.164
	False Positive	3.917
	True Negative	61.583
	False Negative	7.440
Matrik Evaluasi	Akurasi	85,08%
	Presisi	44,68%
	Recall	29,84%
	F1-*Score*	35,78%

Pada skenario model 3, dilakukan penanganan terhadap ketidakseimbangan data menggunakan teknik kombinasi SMOTE-Tomek *Links* sebelum pelatihan model *Random Forest*. SMOTE (Synthetic Minority Over-sampling Technique) berfungsi untuk menambah jumlah sampel pada kelas minoritas dengan cara mensintesis data baru, sedangkan Tomek *Links* digunakan untuk menghapus pasangan sampel dari kelas berbeda yang saling berdekatan, sehingga membantu membersihkan batas keputusan antara kelas. Teknik ini bertujuan tidak hanya

menyeimbangkan distribusi kelas, tetapi juga meningkatkan kejelasan batas klasifikasi.

Model ini menghasilkan akurasi sebesar 85,08%, presisi 44,68%, *recall* 29,84%, dan F1-*score* 35,78%. Jika dibandingkan dengan skenario sebelumnya, terlihat peningkatan signifikan terutama pada nilai *recall* dan F1-*score*. *Recall* meningkat dari 9,72% (pada skenario model 2) menjadi 29,84%, yang menunjukkan bahwa model menjadi lebih baik dalam mengenali pasien dengan risiko diabetes. Meskipun nilai *recall* ini masih tergolong rendah, peningkatannya lebih dari dua kali lipat dibandingkan sebelumnya, mengindikasikan adanya perbaikan dalam mendeteksi kelas minoritas.

4.2.4 Analisis Skenario Model 4

Tabel 4. 11 Analisa skenario model 4

Analisa Skenario Model 4		
Best parameter	N_estimators	300
	Max_depth	None
Confusion Matrix	*True Positive*	6.414
	False Positive	11.459
	True Negative	54.041
	False Negative	4.190
Matrik Evaluasi	Akurasi	79,44%
	Presisi	35,89%
	Recall	60,49%
	F1-*Score*	45,05%

Skenario model 4 mengombinasikan dua pendekatan utama dalam upaya meningkatkan performa klasifikasi, yaitu penyeimbangan data menggunakan SMOTE-Tomek Links dan seleksi fitur menggunakan RFE Pendekatan ini dirancang untuk menangani dua isu utama yang memengaruhi performa model pada skenario sebelumnya: ketidakseimbangan distribusi kelas dan fitur yang tidak relevan dalam data.

Model ini menghasilkan nilai akurasi sebesar 79,44%, presisi 35,89%, *recall* 60,49%, dan F1-*score* 45,05%. Jika dibandingkan dengan model 3 (yang menggunakan SMOTE-Tomek Links saja), terlihat bahwa *recall* meningkat dari 29,84% (model 3) menjadi 60,49%, F1-*score* juga mengalami peningkatan, dari 35,78% menjadi 45,05%, dan Presisi menurun dari 44,68% (model 3) menjadi 35,89%, mengindikasikan adanya peningkatan jumlah *false positive* dari 3.917 menjadi 11.459.

Penurunan presisi ini adalah *trade-off* dari strategi model yang menjadi lebih agresif dalam mengidentifikasi kasus positif setelah seleksi fitur dilakukan. Dengan jumlah *true positive* meningkat (dari 3.164 menjadi 6.414) dan *false negative* menurun (dari 7.440 menjadi 4.190), dapat disimpulkan bahwa model menjadi lebih sensitif terhadap kasus diabetes.

Kombinasi RFE setelah proses penyeimbangan data memperkuat kemampuan model dalam mengenali pola penting pada kelas minoritas. Seleksi fitur membantu mengurangi redundansi dan *noise* yang mungkin masih terdapat pada model 3, memungkinkan model untuk lebih fokus pada fitur-fitur yang benar-benar relevan terhadap diagnosis diabetes.

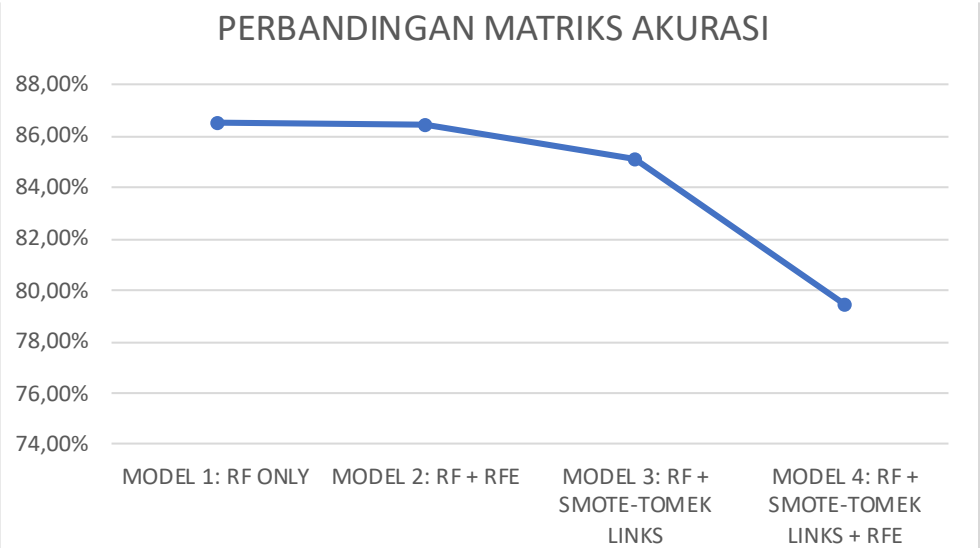
Namun, dampaknya terhadap akurasi menunjukkan sedikit penurunan (dari 85,08% pada model 3 menjadi 79,44%). Ini adalah *trade-off* yang umum terjadi dalam kasus klasifikasi dengan data tidak seimbang, di mana peningkatan sensitivitas terhadap kelas minoritas (*recall*) sering kali dicapai dengan sedikit mengorbankan akurasi keseluruhan. Dalam konteks deteksi medis seperti ini, peningkatan kemampuan mendeteksi pasien berisiko memiliki nilai yang lebih

besar dibandingkan sekadar mempertahankan akurasi global (Pinandito, Wicaksono, & Wijoyo, 2023).

4.3 Analisis Skenario Model Berdasarkan Matriks Evaluasi

Untuk memahami secara mendalam dampak dari pendekatan yang digunakan, perlu dilakukan analisis secara terpisah terhadap empat metrik evaluasi utama, yaitu akurasi, presisi, *recall*, dan f1-*score.*

4.3.1 Akurasi

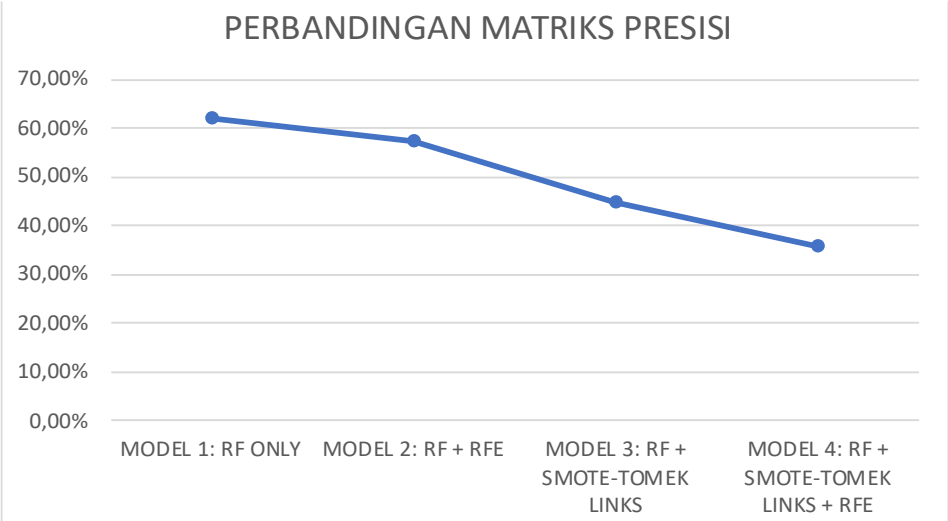


Gambar 4. 8 Perbandingan akurasi model tiap skenario

Gambar 4.8 menunjukkan perbandingan matriks akurasi pada setiap skenario model. Akurasi menunjukkan proporsi prediksi yang benar terhadap seluruh data, baik dari kelas positif maupun negatif. Model 1 dan Model 2 menghasilkan akurasi tertinggi (86,56% dan 86,42%), tetapi hasil ini menyesatkan karena diperoleh dari dominasi prediksi terhadap kelas mayoritas (non-diabetes). Dalam *dataset* tidak seimbang, akurasi cenderung bias dan tidak mencerminkan

performa sesungguhnya terhadap deteksi kasus minoritas. Penurunan akurasi pada Model 3 dan 4 justru mengindikasikan bahwa model mulai memberikan perhatian pada kelas minoritas, sehingga akurasi bukanlah metrik utama yang dijadikan dasar pemilihan model terbaik dalam penelitian ini.

4.3.2 Presisi

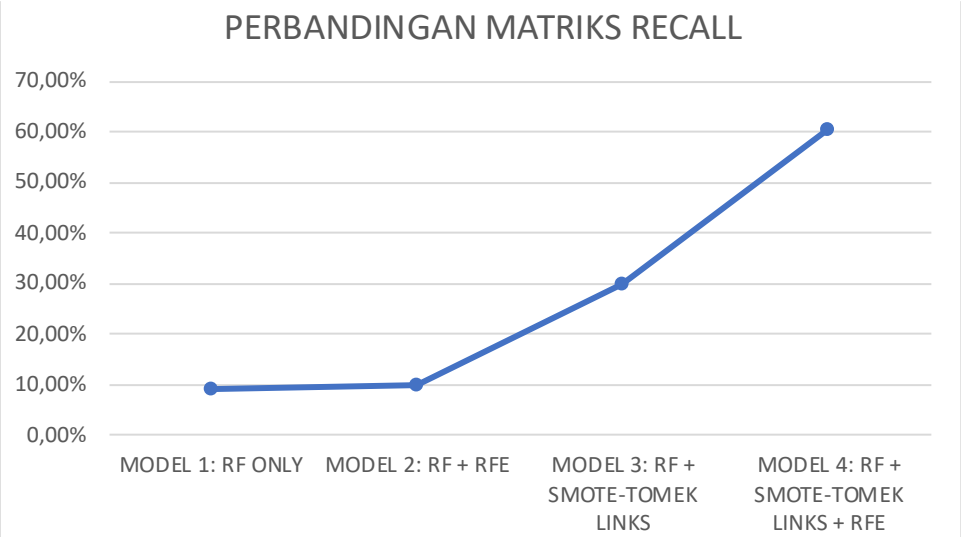


Gambar 4. 9 Perbandingan presisi model tiap skenario

Gambar 4.9 menunjukkan perbandingan matriks presisi pada setiap skenario model. Presisi mengukur sejauh mana prediksi positif yang dihasilkan model benar-benar berasal dari kelas positif. Nilai presisi tertinggi terdapat pada Model 1 (66,42%), yang artinya sebagian besar prediksi positif adalah benar. Namun, jumlah prediksi positif ini sangat kecil karena *recall*-nya rendah. Seiring meningkatnya *recall* dari Model 1 ke Model 4, presisi menurun karena jumlah *false positive* meningkat. Penurunan presisi pada Model 4 ke angka 35,89% merupakan *trade-off* yang umum ketika model menjadi lebih agresif dalam mendeteksi kelas positif. Hal ini masih dapat diterima karena dalam konteks medis, lebih baik mendeteksi lebih

banyak pasien berisiko dengan risiko *false positive*, daripada tidak mendeteksi mereka sama sekali.

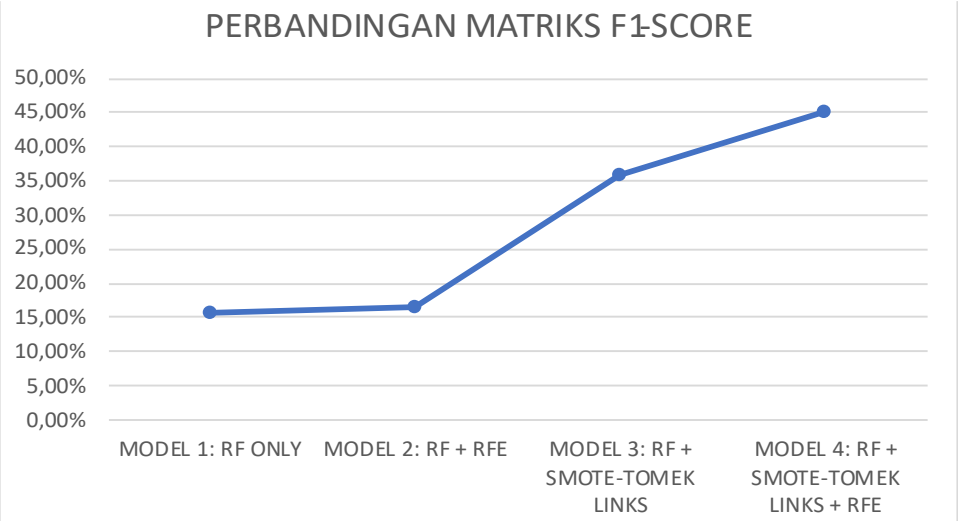
4.3.3 *Recall*



Gambar 4. 10 Perbandingan *recall* model tiap skenario

Gambar 4.10 menunjukkan perbandingan matriks *recall* pada setiap skenario model. *Recall* merupakan metrik kunci dalam penelitian ini. Nilai *recall* yang rendah berarti banyak pasien berisiko tidak terdeteksi. Model 1 dan 2 memiliki nilai *recall* sangat rendah (sekitar 9%), yang menunjukkan ketidakefektifan dalam mengidentifikasi kasus minoritas. Model 3 meningkatkan *recall* menjadi 29,84%, dan Model 4 mencapai *recall* tertinggi sebesar 60,49%. Ini berarti, dari seluruh pasien yang seharusnya terdeteksi sebagai berisiko, 60,49% berhasil dikenali oleh Model 4. Walaupun angka ini masih jauh dari sempurna, peningkatan lebih dari dua kali lipat dibanding model 1 menunjukkan efektivitas kombinasi penyeimbangan data dan seleksi fitur.

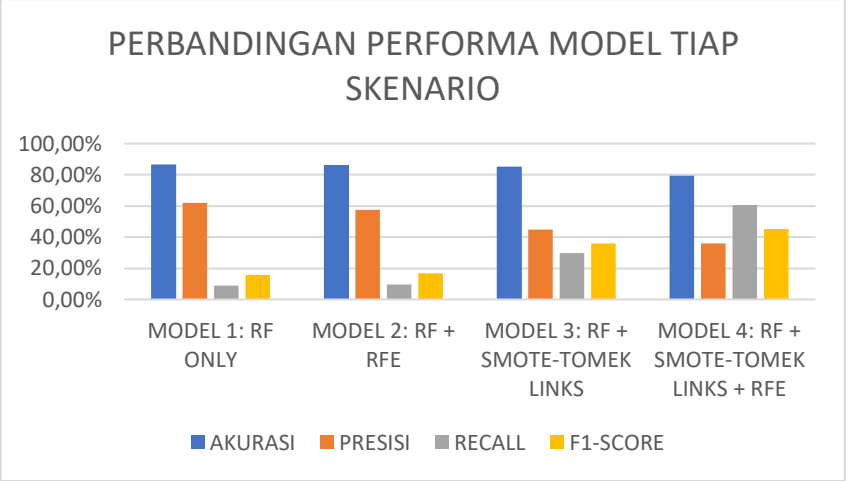
4.3.4 F1-*Score*



Gambar 4. 11 Perbandingan f1-*score* model tiap skenario

Gambar 4.11 menunjukkan perbandingan matriks f1-*score* pada setiap skenario model. F1-*score* mencerminkan keseimbangan antara presisi dan *recall*, dan sangat relevan untuk kasus medis. Model 1 dan model 2 memiliki f1-score yang relatif kecil yaitu 15,73% dan 16,63% sedangkan Model 3 meningkat drastis menjadi 35,78% dan Model 4 mencapai nilai tertinggi yaitu 45,05%. Hal ini membuktikan bahwa kombinasi SMOTE-Tomek *Links* dan RFE berhasil menciptakan model yang tidak hanya sensitif dalam mendeteksi pasien berisiko, tetapi juga cukup presisi sehingga prediksinya tetap dapat dipercaya.

4.4 Kesimpulan Analisis Perbandingan Skenario Model



Gambar 4. 12 Perbandingan performa model tiap skenario

Gambar 4.12 menunjukkan perbandingan performa model pada setiap skenario model. Secara umum, akurasi bukanlah indikator tunggal untuk menilai performa pada data medis yang tidak seimbang. Dalam konteks penelitian ini, metrik seperti *recall* dan f1-*score* jauh lebih penting karena fokus utama adalah mendeteksi pasien yang berisiko diabetes. *Recall* mengukur seberapa baik model dalam mengenali seluruh kasus positif (pasien berisiko), sementara f1-*score* memberikan gambaran keseimbangan antara presisi dan *recall*. Kedua metrik ini sangat relevan digunakan dalam kasus medis di mana kesalahan dalam mendeteksi pasien yang berisiko (*false negative*) dapat berdampak serius terhadap pencegahan dan penanganan penyakit.

Tabel 4. 12 Performa model pada setiap skenario

Skenario	Akurasi	Presisi	Recall	F1-Score
Model 1: Tanpa Penanganan Data & Seleksi Fitur	86,56%	62,13%	9,01%	15,73%
Model 2: Seleksi Fitur RFE	86,42%	57,47%	9,72%	16,63%
Model 3: Penanganan Data SMOTE-Tomek *Links*	85,08%	44,68%	29,84%	35,78%
Model 4: SMOTE-Tomek *Links* + RFE	79,44%	35,89%	60,49%	45,05%

Berdasarkan tabel 4.12 di atas, dapat terlihat bahwa setiap skenario yang diterapkan memberikan dampak berbeda pada performa model. Model 1 yang menggunakan data asli tanpa penanganan *imbalance* maupun seleksi fitur menghasilkan akurasi tertinggi sebesar 86,56% dan presisi tertinggi sebesar 66,42%. Namun, model ini hanya memperoleh nilai *recall* sebesar 9,01%, yang artinya model gagal mendeteksi sebagian besar pasien yang sebenarnya berisiko diabetes. Model 2 yang menerapkan seleksi fitur menggunakan RFE menunjukkan sedikit peningkatan pada *recall* menjadi 9,72% dan f1-*score* tetap di angka 16,14%. Hal ini menunjukkan bahwa permasalahan utama bukan terletak pada jumlah fitur, tetapi pada distribusi kelas yang tidak seimbang.

Perubahan performa model yang lebih signifikan mulai terlihat pada Model 3 yang menerapkan teknik penanganan ketidakseimbangan data menggunakan SMOTE-Tomek *Links*. Dengan metode ini, *recall* meningkat menjadi 29,84% dan f1-*score* naik menjadi 35,78%, meskipun akurasi mengalami sedikit penurunan menjadi 85,08%. Peningkatan ini menunjukkan bahwa penyeimbangan data memiliki pengaruh yang besar terhadap kemampuan model dalam mengenali kelas minoritas (pasien berisiko), dibandingkan hanya melakukan seleksi fitur.

Model 4 merupakan skenario yang menggabungkan dua pendekatan sekaligus, yaitu SMOTE-Tomek *Links* dan seleksi fitur RFE. Hasil yang diperoleh menunjukkan bahwa model ini memiliki performa terbaik dalam mendeteksi pasien berisiko, dengan *recall* tertinggi sebesar 60,49% dan f1-*score* sebesar 45,05%. Meskipun akurasi dan presisi mengalami penurunan, ini merupakan *trade-off* yang

dapat diterima dalam konteks klasifikasi medis. Keberhasilan dalam mendeteksi pasien yang benar-benar berisiko lebih diutamakan daripada akurasi umum pada data yang tidak seimbang.

Dengan demikian, model 4 dinilai paling seimbang dan relevan untuk digunakan sebagai model klasifikasi risiko diabetes karena berhasil menggabungkan keakuratan prediksi dengan efisiensi dan sensitivitas terhadap data kelas minoritas.

4.5 Integrasi Sains dan Islam

Pengembangan model klasifikasi risiko diabetes, tidak hanya bermanfaat dari sisi akademis, tetapi juga memiliki nilai yang sejalan dengan ajaran Islam. Dalam Islam, pemanfaatan ilmu pengetahuan untuk kebaikan merupakan bentuk rasa syukur kepada Allah yang telah menciptakan alam semesta beserta isinya untuk kemaslahatan umat manusia. Hal ini ditegaskan dalam Al-Qur'an:

وَسَخَّرَ لَكُمْ مَّا فِى السَّمٰوٰتِ وَمَا فِى الْاَرْضِ جَمِيْعًا مِّنْهُۗ اِنَّ فِيْ ذٰلِكَ لَاٰيٰتٍ لِّقَوْمٍ يَّتَفَكَّرُوْنَ

"Dan Dia telah menundukkan untukmu apa yang di langit dan apa yang di bumi semuanya, (sebagai rahmat) daripada-Nya. Sesungguhnya pada yang demikian itu benar-benar terdapat tanda-tanda (kekuasaan Allah) bagi kaum yang berpikir." (Q.S Al Jatsiyah : 13).

Dalam Tafsir Jalalain dijelaskan bahwa yang dimaksud *apa yang ada di langit* mencakup matahari, bulan, bintang-bintang, air hujan, dan lain-lainnya; sedangkan *apa yang ada di bumi* mencakup binatang, pohon-pohon, tumbuh-tumbuhan, sungai, dan sebagainya. Semuanya diciptakan oleh Allah untuk dimanfaatkan oleh manusia. Kata *jami'an* (semuanya) berfungsi sebagai penegasan, dan *minhu* (dari-Nya) menunjukkan bahwa semua itu tunduk kepada

kekuasaan-Nya (TafsirQ, t.t.-a). Ayat ini menjadi tanda-tanda kekuasaan dan keesaan Allah bagi orang-orang yang mau berpikir, sehingga mereka beriman. Dengan demikian, pemanfaatan ilmu pengetahuan dalam bidang kesehatan, seperti pada penelitian ini, adalah wujud nyata pemanfaatan anugerah Allah dengan bijak untuk kesejahteraan umat.

Selain sebagai bentuk syukur kepada Allah, pemanfaatan ilmu pengetahuan dalam bidang kesehatan juga mencerminkan kepedulian dan tolong-menolong antar sesama. Islam sangat menekankan pentingnya saling membantu dalam kebaikan dan takwa, sebagaimana disebutkan dalam penggalan QS. Al Maidah ayat 2 berikut:

وَتَعَاوَنُوا عَلَى الْبِرِّ وَالتَّقْوَىٰ وَلَا تَعَاوَنُوا عَلَى الْإِثْمِ وَالْعُدْوَانِ

"Dan tolong-menolonglah kamu dalam (mengerjakan) kebajikan dan takwa, dan jangan tolong-menolong dalam berbuat dosa dan permusuhan." (QS. Al-Maidah: 2).

Dalam Tafsir Jalalain, dijelaskan bahwa perintah "tolong-menolonglah kamu dalam kebaikan" berarti bekerja sama dalam mengerjakan apa yang diperintahkan Allah, sedangkan "dalam ketakwaan" berarti saling membantu untuk meninggalkan apa yang dilarang Allah. Sementara larangan "janganlah kamu bertolong-tolongan dalam berbuat dosa" maksudnya adalah larangan untuk saling mendukung dalam kemaksiatan, dan "pelanggaran" artinya melampaui batas-batas yang telah ditetapkan Allah. Ayat ini diakhiri dengan perintah untuk bertakwa, yaitu takut kepada azab-Nya dengan menaati perintah-Nya, karena sesungguhnya Allah sangat berat siksa-Nya bagi mereka yang melanggar (TafsirQ, t.t.-b).

Ayat dan tafsir ini menjadi penguat penting bahwa usaha-usaha dalam bidang kesehatan, termasuk pengembangan model klasifikasi risiko diabetes, merupakan bagian dari bentuk tolong-menolong dalam kebaikan dan ketakwaan. Dengan membantu mendeteksi risiko penyakit sejak dini, penelitian ini tidak hanya memenuhi tanggung jawab akademis, tetapi juga menjadi wujud nyata kontribusi sosial dalam mencegah *mudarat* (bahaya) bagi sesama manusia.

Islam juga mengajarkan bahwa segala bentuk usaha manusia, termasuk dalam menjaga kesehatan, pada dasarnya adalah bagian dari penghambaan kepada Allah SWT. Upaya mencari pengobatan dan mencegah penyakit bukan hanya demi kepentingan duniawi, tetapi juga bentuk ketaatan dan pengakuan terhadap nikmat Allah SWT. Hal ini sesuai dengan sabda Rasulullah ﷺ:

إِنَّ اللَّهَ أَنْزَلَ الدَّاءَ وَالدَّوَاءَ، وَجَعَلَ لِكُلِّ دَاءٍ دَوَاءً فَتَدَاوَوْا وَلَا تَدَاوَوْا بِحَرَامٍ

"Sesungguhnya Allah menurunkan penyakit dan obatnya dan menjadikan bagi setiap penyakit ada obatnya. Maka berobatlah kalian, dan jangan kalian berobat dengan yang haram." (HR. Abu Dawud dari Abu Darda).

Mayoritas ulama dari kalangan Hanafiyah dan Malikiyah berpendapat bahwa hukum berobat adalah *mubah* atau diperbolehkan, sedangkan ulama Syafi'iyah, Al-Qadhi, Ibnu Aqil, dan Ibnul Jauzi dari kalangan Hambali berpendapat bahwa hukumnya *mustahab* atau dianjurkan (KuttabRuQu, 2021). Mereka berdalil dengan hadist-hadist yang memerintahkan untuk berobat, serta teladan Nabi ﷺ yang melakukan berobat dan bekam sebagai bentuk syariat menjaga kesehatan.

Menurut ulama Syafi'iyah, jika pengobatan diyakini membawa manfaat yang sangat besar, seperti membalut luka atau transfusi darah dalam kondisi tertentu, maka hukumnya bahkan bisa berubah menjadi wajib. Pandangan ini sejalan dengan prinsip *maqashid syariah* yang menekankan pentingnya *hifzh an-nafs* (menjaga jiwa) sebagai salah satu tujuan utama syariat. Upaya preventif melalui deteksi dini risiko diabetes adalah bagian dari implementasi nilai *hifzh an-nafs,* karena dapat mengurangi risiko komplikasi serius bahkan kematian akibat diabetes.

Ibnul Qayyim dalam *Zadul Ma'ad* menekankan bahwa berobat tidak menafikan konsep tawakal. Menurutnya, justru dengan melakukan ikhtiar seperti makan saat lapar atau minum saat haus, seseorang dapat mencapai tawakal yang sempurna. Ibnul Qayyim mengingatkan bahwa meninggalkan ikhtiar bukanlah bentuk ketakwaan, tetapi justru bentuk kelemahan yang bertentangan dengan makna tawakal sejati. Tawakal yang benar adalah mengandalkan hati kepada Allah sambil tetap melakukan usaha lahiriah yang telah Allah tetapkan sebagai sebab terjadinya takdir (Al Manhaj, 2010).

Dalam konteks penelitian ini, pengembangan model klasifikasi risiko diabetes dapat dilihat sebagai bagian dari ikhtiar yang diperintahkan dalam syariat. Penelitian ini berperan membantu upaya pencegahan dan deteksi dini penyakit, yang bukan hanya bermanfaat secara akademis, tetapi juga bernilai ibadah karena membantu menjaga kesehatan umat.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini bertujuan untuk mengimplementasikan algoritma *Random Forest* dengan kombinasi SMOTE-TOMEK *Links* dan *Recursive Feature Elimination (RFE)* serta menganalisis performa model dalam klasifikasi risiko diabetes. Berdasarkan hasil penelitian yang telah dilakukan mengenai Klasifikasi Risiko Diabetes menggunakan *Random Forest* dengan Kombinasi SMOTE-Tomek *Links* dan *Recursive Feature Elimination*, dapat disimpulkan bahwa Penelitian ini berhasil mengimplementasikan algoritma *Random Forest* dalam empat skenario model klasifikasi risiko diabetes, yaitu tanpa penanganan data dan seleksi fitur, dengan seleksi fitur RFE saja, dengan SMOTE-Tomek *Links* saja, dan gabungan SMOTE-Tomek *Links* dan RFE. Teknik SMOTE-Tomek *Links* digunakan untuk menangani ketidakseimbangan kelas pada *dataset*, sedangkan RFE digunakan untuk memilih fitur yang paling relevan guna meningkatkan efisiensi model.

Hasil evaluasi menunjukkan bahwa model terbaik adalah model 4, yaitu model yang menerapkan metode penyeimbangan data SMOTE-Tomek *Links* dan seleksi fitur RFE. Meskipun akurasi dan presisi model ini mengalami penurunan dibanding model awal (Model 1), *recall* dan f1-*score* justru mengalami peningkatan signifikan. *Recall* meningkat dari 9,01% (Model 1) menjadi 60,49% (Model 4), dan f1-*score* meningkat dari 16,14% menjadi 45,05%. Hal ini menandakan bahwa model 4 memiliki kemampuan yang jauh lebih baik dalam mendeteksi pasien yang berisiko diabetes, yang menjadi fokus utama dalam klasifikasi berbasis data medis.

Penurunan akurasi dan presisi merupakan *trade-off* yang dapat diterima dalam konteks kesehatan, karena keberhasilan dalam mengidentifikasi pasien berisiko lebih diutamakan daripada sekedar menjaga nilai akurasi.

5.2 Saran

Dari hasil penelitian dan proses uji coba yang telah dilakukan, peneliti menyadari bahwa penelitian ini masih banyak keterbatasan dan kekurangan. Oleh karena itu peneliti memberikan beberapa saran untuk pengembangan lebih lanjut:

1. Disarankan untuk mempertimbangkan perbandingan performa antara algoritma *Random Forest* dengan algoritma lain seperti DNN, KNN, dan SVM untuk mengetahui pendekatan mana yang paling optimal dalam klasifikasi risiko diabetes.
2. Disarankan untuk mempertimbangkan perbandingan performa antara metode penyeimbangan data SMOTE-Tomek *Links* dengan metode lain seperti SMOTE-ENN, dan ADASYN untuk mengetahui pendekatan mana yang paling optimal dalam penanganan *dataset* yang tidak seimbang pada klasifikasi risiko diabetes.
3. Disarankan untuk melakukan verifikasi fitur yang dihasilkan RFE dengan orang yang memiliki kapasitas di bidang kesehatan sebelum digunakan sebagai model klasifikasi risiko diabetes.

DAFTAR PUSTAKA

Abdulhadi, N., & Al-Mousa, A. (2021). Diabetes Detection Using Machine Learning Classification Methods. *2021 International Conference on Information Technology (ICIT)*, 350–354. Amman, Jordan: IEEE. https://doi.org/10.1109/ICIT52682.2021.9491788

Al Manhaj. (2010, Maret 29). Hukum Berobat. Diambil 9 Mei 2025, dari https://almanhaj.or.id/2691-hukum-berobat.html

Alva, M., Gray, A., Mihaylova, B., & Clarke, P. (2014). THE EFFECT OF DIABETES COMPLICATIONS ON HEALTH-RELATED QUALITY OF LIFE: THE IMPORTANCE OF LONGITUDINAL DATA TO ADDRESS PATIENT HETEROGENEITY. *Health Economics*, *23*(4), 487–500. https://doi.org/10.1002/hec.2930

American Diabetes Association. (2020). *Standards of Medical Care in Diabetes—2020* Abridged for Primary Care Providers. *Clinical Diabetes*, *38*(1), 10–38. https://doi.org/10.2337/cd20-as01

Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4302-5990-9

Castelli, M., Vanneschi, L., & Largo, Á. R. (2019). Supervised Learning: Classification. Dalam *Encyclopedia of Bioinformatics and Computational Biology* (hlm. 342–349). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20332-4

Chen, S. (2023). Comparison of machine learning algorithms and feature visualization analysis for diabetes risk prediction. *Journal of Physics: Conference Series*, *2646*(1), 012013. https://doi.org/10.1088/1742-6596/2646/1/012013

Chowdhury, M. M., Ayon, R. S., & Hossain, M. S. (2024). An investigation of machine learning algorithms and data augmentation techniques for diabetes diagnosis using class imbalanced BRFSS dataset. *Healthcare Analytics*, *5*, 100297. https://doi.org/10.1016/j.health.2023.100297

Devia, A., & Soewito, B. (2023). *Analisis Perbandingan Metode Seleksi Fitur untuk Mendeteksi Anomali pada Dataset CIC-IDS-2018*. *5*(4).

Dewi, P., Azizah, M., Rendowaty, A., Sri Wahyuni, Y., & Pranata, L. (2023). Edukasi tentang Diabetes Mellitus dan Pemeriksaan Biomedis Kadar Gula Darah Pada Ibu Rumah Tangga. *Health Community Service*, *1*(1), 46–50. https://doi.org/10.47709/hcs.v1i1.3358

Dhani, A. A., Siswa, T. A. Y., & Pranoto, W. J. (2024). Perbaikan Akurasi Random Forest Dengan ANOVA Dan SMOTE Pada Klasifikasi Data Stunting. *Teknika*, *13*(2), 264–272. https://doi.org/10.34148/teknika.v13i2.875

Herman, W. H., Ye, W., Griffin, S. J., Simmons, R. K., Davies, M. J., Khunti, K., … Wareham, N. J. (2015). Early Detection and Treatment of Type 2 Diabetes Reduce Cardiovascular Morbidity and Mortality: A Simulation of the Results of the Anglo-Danish-Dutch Study of Intensive Treatment in People With Screen-Detected Diabetes in Primary Care (ADDITION-Europe). *Diabetes Care*, *38*(8), 1449–1455. https://doi.org/10.2337/dc14-2459

Ibrahim, M. (2023). Evolution of Random Forest from Decision Tree and Bagging: A Bias-Variance Perspective. *Dhaka University Journal of Applied Science and Engineering*, *7*(1), 66–71. https://doi.org/10.3329/dujase.v7i1.62888

Iskandar, R. F. N., Gutama, D. H., Wijaya, D. P., & Danianti, D. (2024). Klasifikasi Menggunakan Metode Random Forest untuk Awal Deteksi Diabetes Melitus Tipe 2. *Jurnal Teknik Industri Terintegrasi*, *7*(3), 1620–1626. https://doi.org/10.31004/jutin.v7i3.26916

Ismail, L., Materwala, H., & Al Kaabi, J. (2021). Association of risk factors with type 2 diabetes: A systematic review. *Computational and Structural Biotechnology Journal*, *19*, 1759–1785. https://doi.org/10.1016/j.csbj.2021.03.003

Kumar, P., Bhatnagar, R., Gaur, K., & Bhatnagar, A. (2021). Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*, *1099*(1), 012077. https://doi.org/10.1088/1757-899X/1099/1/012077

KuttabRuQu. (2021, Agustus 22). Bagaimana Hukum Berobat dengan Suatu Obat? -. Diambil 9 Mei 2025, dari https://kuttab-rumahquran.com/index.php/2021/08/22/bagaimana-hukum-berobat-dengan-suatu-obat/

Mulyani, E., Muhamad, F. P. B., & Cahyanto, K. A. (2021). Pengaruh N-Gram terhadap Klasifikasi Buku menggunakan Ekstraksi dan Seleksi Fitur pada Multinomial Naïve Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(1), 264. https://doi.org/10.30865/mib.v5i1.2672

Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, & Susanti. (2022). Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest. *SATIN - Sains dan Teknologi Informasi*, *8*(2), 122–131. https://doi.org/10.33372/stn.v8i2.885

Ozougwu, O. (2013). The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. *Journal of Physiology and Pathophysiology*, *4*(4), 46–57. https://doi.org/10.5897/JPAP2013.0001

Patel, H., Singh Rajput, D., Thippa Reddy, G., Iwendi, C., Kashif Bashir, A., & Jo, O. (2020). A review on classification of imbalanced data for wireless sensor networks. *International Journal of Distributed Sensor Networks*, *16*(4), 155014772091640. https://doi.org/10.1177/1550147720916404

Pierannunzi, C., Hu, S. S., & Balluz, L. (2013). A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004–2011. *BMC Medical Research Methodology*, *13*(1), 49. https://doi.org/10.1186/1471-2288-13-49

Pratama, A. R. I., Latipah, S. A., & Sari, B. N. (2022). OPTIMASI KLASIFIKASI CURAH HUJAN MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) DAN RECURSIVE FEATURE ELIMINATION (RFE). *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *7*(2), 314–324. https://doi.org/10.29100/jipi.v7i2.2675

Rasheed, S., Kiran Kumar, G., Rani, D. M., Prasad Kantipudi, M. V. V., & M, A. (2024). Heart Disease Prediction Using GridSearchCV and Random Forest. *EAI Endorsed Transactions on Pervasive Health and Technology*, *10*. https://doi.org/10.4108/eetpht.10.5523

Rezki, M. K., Mazdadi, M. I., Indriani, F., Saragih, H., & Athavale, V. A. (2024). *Application of Smote to Address Class Imbalance in Diabetes Disease Categorization Utilizing C5.0, Random Forest, and Support Vector Machine*. *6*(4).

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *5*(1). https://doi.org/10.31294/ijcit.v5i1.7951

Satria, B., Siswa, T. A. Y., & Pranoto, W. J. (2024). Optimasi Random Forest dengan Genetic Algorithm dan Recursive Feature Elimination pada High Dimensional Data Stunting Samarinda. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *8*(3), 1778. https://doi.org/10.30865/mib.v8i3.7883

Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4). https://doi.org/10.1214/15-AOS1321

Septiana Rizky, P., Haiban Hirzi, R., & Hidayaturrohman, U. (2022). Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang. *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, *15*(2), 228–236. https://doi.org/10.36456/jstat.vol15.no2.a5548

Shabrina Assyifa, D., & Luthfiarta, A. (2024). SMOTE-Tomek Re-sampling Based on Random Forest Method to Overcome Unbalanced Data for Multi-class Classification. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(2), 151–160. https://doi.org/10.25139/inform.v9i2.8410

Sir, Y. A., & Soepranoto, A. H. H. (2022). Pendekatan Resampling Data Untuk Menangani Masalah Ketidakseimbangan Kelas. *Jurnal Komputer dan Informatika*, *10*(1), 31–38. https://doi.org/10.35508/jicon.v10i1.6554

Suratun, S., Pujiana, D., & Sari, M. (2023). PENCEGAHAN DIABETES MELITUS DI PALEMBANG. *Masker Medika*, *11*(1), 9–18. https://doi.org/10.52523/maskermedika.v11i1.507

Swana, E. F., Doorsamy, W., & Bokoro, P. (2022). Tomek Link and SMOTE Approaches for Machine Fault Classification with an Imbalanced Dataset. *Sensors*, *22*(9), 3246. https://doi.org/10.3390/s22093246

TafsirQ. (t.t.-a). Surat Al-Jasiyah Ayat 13. Diambil 9 Mei 2025, dari TafsirQ website: https://tafsirq.com/permalink/ayat/4486

TafsirQ. (t.t.-b). Surat Al-Ma'idah Ayat 2. Diambil 10 Mei 2025, dari https://tafsirq.com/permalink/ayat/671

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/2557795

Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., … Qiu, L. (2021). Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Medical Informatics and Decision Making*, *21*(1), 105. https://doi.org/10.1186/s12911-021-01471-4

Zhou, H., Xin, Y., & Li, S. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. *BMC Bioinformatics*, *24*(1), 224. https://doi.org/10.1186/s12859-023-05300-5