KLASIFIKASI PEROKOK BERDASARKAN KONDISI TUBUH MENGGUNAKAN METODE *RANDOM FOREST*

SKRIPSI

Oleh:
MILA AMARILA PRAMESWARI
NIM. 200605110080

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2024

KLASIFIKASI PEROKOK BERDASARKAN KONDISI TUBUH MENGGUNAKAN METODE *RANDOM FOREST*

SKRIPSI

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

Oleh:
MILA AMARILA PRAMESWARI
NIM. 200605110080

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2024

HALAMAN PERSETUJUAN

KLASIFIKASI PEROKOK BERDASARKAN KONDISI TUBUH MENGGUNAKANMETODE *RANDOM FOREST*

SKRIPSI

Oleh:
MILA AMARILA PRAMESWARI
NIM. 200605110080

Telah Diperiksa dan Disetujui untuk Diuji
Tanggal: 22 Mei 2024

Pembimbing I	Pembimbing II
Prof. Dr. Suhartono, 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. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

KLASIFIKASI PEROKOK BERDASARKAN KONDISI TUBUH MENGGUNAKAN METODE *RANDOM FOREST*

SKRIPSI

Oleh:
MILA AMARILA PRAMESWARI
NIM. 200605110080

Telah dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 07 Juni 2024

Susunan Dewan Penguji

Ketua Penguji	: Hani Nurhayati, M.T NIP. 19780625 200801 2 006	( )
Anggota Penguji I	: Tri Mukti Lestari, M.Kom NIP. 19911108 202012 2 005	( )
Anggota Penguji II	: Prof. Dr. Suhartono, 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. Fachrul Kurniawan, M.MT, IPM
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Mila Amarila Prameswari
NIM	: 200605110080
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Klasifikasi Perokok Berdasarkan Kondisi Tubuh Menggunakan Metode *Random Forest*

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, 07 Juni 2024
Yang membuat pernyataan,

Mila Amarila Prameswari
NIM. 200605110080

HALAMAN MOTTO

“Boleh jadi kamu membenci sesuatu, padahal ia amat baik bagimu, dan boleh jadi pula kamu menyukai sesuatu, padahal ia amat buruk bagimu. Allah mengetahui, sedang kamu tidak mengetahui”

(Q.S. Al-Baqarah: 216)

“Tidak peduli seberapa sulit atau mustahilnya itu, jangan pernah melupakan tujuanmu”

-Luffy-

HALAMAN PERSEMBAHAN

Alhamdulillah Wasyukurillah puji syukur kehadirat Allah SWT. Atas limpahan Rahmat, Taufiq, dan Hidayah kepada penulis Serta sholawat serta salam bagi Rasulullah SAW yang telah membawa cahaya Islam dan teladan yang sempurna bagi umat manusia. Dengan rasa hormat dan terima kasih, penulis mempersembahkan skripsi tugas akhir ini kepada:

1. Ibu saya tercinta Ibu Atik, yang tak pernah lelah memberikan cinta, dukungan dan doa serta kasih sayang dalam setiap langkah hidup penulis. Beribu-ribu terima kasih penulis sampaikan karena telah menjadi sumber inspirasi yang tak ternilai, cahaya yang menerangi jalan dan kekuatan yang menguatkan hati penulis.
2. Ayah saya tercinta Ayah Imam, yang selalu menjadi teladan dalam hidup penulis dengan segenap perjuangan yang telah beliau lakukan, serta kasih sayang yang disampaikan dalam cara yang berbeda.
3. Adik saya tersayang Nuchy yang tak pernah lelah membuat penulis terheran-heran, yang membuat penulis kesal namun tak pernah lepas dari rasa sayangmu. Terima kasih untuk semua tawa, dan momen tak terlupakan bersama.
4. Bapak Prof. Dr. Suhartono, M.Kom dan Bapak Dr. Irwan Budi Santoso, M.Kom selaku Dosen Pembimbing Penulis yang sabar serta tulus dalam memberikan arahan dan bimbingan kepada penulis dalam proses penyelesaian tugas akhir.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Puji dan syukur atas kehadirat Tuhan Yang Maha Esa, Allah Subhanahu Wa Ta'ala yang telah memberikan Taufik dan Hidayah-Nya kepada penulis sehingga dapat penyelesaikan skripsi yang berjudul "Klasifikasi Perokok Berdasarkan Kondisi Tubuh Menggunakan Metode Random Forest" dengan baik.

Dalam penulisan skripsi ini banyak pihak yang terlibat dalam bentuk doa, restu, dan bimbingan dari berbagai pihak serta niat penulis sendiri. Oleh sebab itu, dalam kesempatan kali ini penulis ingin mengucapkan banyak terima kasih kepada:

1. Prof. Dr. H. M. Zainuddin, MA, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang beserta jajarannya.
2. Prof. Dr. Sri Harini, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang beserta jajarannya.
3. Dr. Fachrul Kurniawan ST., M.MT, IPM selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Prof. Dr. Suhartono, M.Kom selaku Dosen Pembimbing I yang telah dengan sabar dalam memberikan arahan dalam penulisan hingga program yang dibuat dalam menyelesaikan skripsi ini.
5. Dr. Irwan Budi Santoso, M.Kom selaku Dosen Pembimbing II yang telah membimbing penlis dalam menyelesaikan skripsi ini.
6. Dr. Cahyo Crysdian, M.Cs selaku Dosen Wali yang telah memberikan arahan dalam proses perkuliahan.
7. Segenap Dosen, Laboran, dan jajaran pada Program Studi Teknik Informatika yang telah memberikan bimbingan dan bantuan selama studi.
8. Nia Faricha, S.Si selaku Admin Program Studi Teknik Informatika yang dengan sabar membantu, memberikan arahan, informasi terkait perkuliahan.
9. Kedua orang tua penulis Ibu Atik dan Ayah Imam yang selalu memberikan dukungan dan perhatian serta selalu mendoakan yang terbaik untuk kelancaran proses pendidikan putrinya.
10. Adik penulis Nuchy yang menemani dan menghibur penulis.

11. Sahabat penulis “Gantadrie” yang beranggotakan Kartika dan Nadia terima kasih atas semua bantuan dan semangat yang telah kalian berikan dan kenangan dari awal kuliah tatap muka pertama hingga saat ini, *see you on top guys*.
12. Teman dekat penulis yaitu Alfina, Ina, Rizka, Vera, Niken, dan Firoh. Terima kasih telah banyak membantu penulis selama masa perkuliahan hingga saat ini.
13. Seluruh keluarga besar Saudara Teknik Informatika UIN Malang terkhusus Angkatan 2020 “Integer”. serta teman-teman yang dekat dengan penulis tidak bisa disebutkan satu persatu, terimakasih telah memberikan *support*, motivasi dan bantuannya kepada penulis.
14. Seluruh pihak yang telah terlibat secara langsung maupun tidak langsung dalam proses penyusunan skripsi sejauh ini.
15. Seluruh member “TOMORROW X TOGETHER” khususnya BEOMGYU yang telah menghibur dan memberikan semangat secara tidak langsung kepada penulis. Serta seluruh kru SHP yang hadir saat penulis sedang dalam keadaan *down*.
16. Diri penulis sendiri, terima kasih telah berusaha bangkit dari rasa yang menahan semangatmu hingga semangat itu membara kembali. Tetaplah berjuang meskipun banyak batu sandungan, jangan pernah berpikir untuk menyerah serta tetap untuk rendah hati dan selalu bersyukur atas nikmat yang telah diberikan Allah SWT.

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

Malang, 07 Juni 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ABSTRAK

Prameswari, Mila Amarila, 2024. ***Klasifikasi Perokok Berdasarkan Kondisi Tubuh Menggunakan Metode Random Forest***. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi. Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M. Kom. (II) Dr. Irwan Budi Santoso, M. Kom.

Kata Kunci: *Klasifikasi, Perokok, Random Forest*

Perkembangan teknologi mempengaruhi pola hidup manusia dan meningkatkan resiko berbagai penyakit, termasuk akibat pola hidup tidak sehat dan merokok yang dapat menyebabkan penyakit yang berpotensi meningkatkan risiko kematian. Dalam penelitian ini digunakan metode Random Forest untuk mengklasifikasikan perokok. Tujuan penelitian ini adalah mengetahui *Performance* metode *Random Forest* melalui perhitungan accuracy, precision, recall dan f1-score dalam mengklasifikasikan perokok berdasarkan kondisi tubuh dengan menerapkan *Random Forest*. Data yang digunakakn dalam penelitian ini adalah data "*Body Signal Of Smoking*" yang diambil dari salah satu Perusahaan Asuransi Kesehatan Nasional milik Republik Korea Selatan**.** Kemudian dilakukan preprocessing data meliputi seleksi fitur, *SMOTE*, split data dan normalisasi. Terdapat dua model split data, Model A dengan data train 80% dan data test 20% serta Model B dengan data train 70% dan data test 30%. Berikutnya pada setiap model dilakukan pengujian yang berbeda yaitu tanpa normalisasi dan normalisasi, tanpa *SMOTE* dan menggunakan *SMOTE*, serta tanpa seleksi fitur dan menggunakan seleksi fitur. Nilai akurasi terbaik terjadi pada pengujian dengan normalisasi menggunakan *SMOTE* dan seleksi fitur dengan hasil nilai nilai *Accuracy* 82,50%, *precision* 81,58%, *recall* 89,39% dan *f1-score* 85,31% pada Model A dan nilai *Accuracy* 81,65%, *precision* 80,73%, *recall* 89,06%, dan f*1-score* 84,69% pada Model B.

ABSTRACT

Prameswari, Mila Amarila, 2024. **Classification of Smokers Based on Body Condition Using Random Forest Method**. Thesis. Informatics Engineering Study Program, Faculty of Science and Technology. State Islamic University Maulana Malik Ibrahim Malang. Advisors: (I) Prof. Dr. Suhartono, M. Kom. (II) Dr. Irwan Budi Santoso, M. Kom.

The development of technology affects human life patterns and increases the risk of various diseases, including due to unhealthy lifestyles and smoking which can cause diseases that have the potential to increase the risk of death. In this study, the Random Forest method was used to classify smokers. The purpose of this study is to determine the *Performance* of the Random Forest method through the calculation of accuracy, precision, recall and f1-score in classifying smokers based on body condition by applying Random Forest. The data used in this study is the "Body Signal Of Smoking" data taken from one of the National Health Insurance Companies owned by the Republic of South Korea. Then data preprocessing is carried out including feature selection, *SMOTE*, split data and normalization. There are two data split models, Model A with 80% train data and 20% test data and Model B with 70% train data and 30% test data. Next, each model is tested differently, namely without normalization and normalization, without SMOTE and using SMOTE, and without feature selection and using feature selection. The best accuracy value occurs in testing with normalization using SMOTE and feature selection with the results of the Accuracy value of 82.50%, precision 81.58%, recall 89.39% and f1-score 85.31% in Model A and Accuracy value 81.65%, precision 80.73%, recall 89.06%, and f1-score 84.69% in Model B.

Keywords: Classification, Smoker, Random Forest

البحث مستخلص

برامسواري، ميلا أماريلا، 2024. تصنيف المدخنين بناءً على حالة الجسم باستخدام طريقة ***Random Forest*** . الأطروحة. برنامج دراسة هندسة المعلوماتية، كلية العلوم والتكنولوجيا. جامعة مولانا مالك إبراهيم مالانج الإسلامية الحكومية. المشرف: (الأول) بروفيسور الدكتور سوهارتونو، ماجستير في الحوسبة. (الثاني)الدكتور إروان بودي سانتوسو، ماجستير في الحوسبة

الكلمات المفتاحية: التصنيف، المدخن، ***Random Forest***

تطوير التكنولوجيا التأثير نمط الحياة الإنسان ويزيد من خطر أمراض مختلفة, بما في ذلك نتيجة نمط الحياة غير الصحي والتدخين يمكن أن يسبب الأمراض التي يحتمل أن تزيد من خطر الوفاة. في هذه الدراسة باستخدام الطريقة ***Random Forest*** للقيام بما يلي تصنيف المدخنين. أهداف هذه الدراسة هي اكتشف أداء الطريقة ***Random Forest*** من خلال الحساب الدقة والدقة والاستدعاء دان f1- نقاط للقيام بما يلي تصنيف المدخنين بناءً على حالة الجسم من خلال تطبيق الطريقة. مجموعة البيانات المستخدمة في هذه الدراسة هي مجموعة البيانات"" " المأخوذة من إحدى شركات التأمين الصحي الوطنية المملوكة لجمهورية كوريا الجنوبية. ثم يتم ذلك المعالجة المسبقة للبيانات بما في ذلك اختيار الميزات, تقنية أخذ العينات الزائدة من الأقليات الاصطناعية تقسيم البيانات وتطبيعها. هناك نوعان من نماذج البيانات المنقسمة المستخدمة في هذه الدراسة أي النموذج A مع بيانات تدريب بنسبة 80% وبيانات اختبار بنسبة 20%، والنموذج B مع بيانات تدريب بنسبة 70% وبيانات اختبار بنسبة 30%. بعد ذلك، تم اختبار كل نموذج بشكل مختلف، أي بدون تطبيع وتطبيع، وبدون ***SMOTE*** وباستخدام ***SMOTE***، وبدون اختيار الميزة وباستخدام اختيار الميزة. تظهر أفضل قيمة دقة في الاختبار مع التطبيع باستخدام ***SMOTE*** واختيار السمات مع نتائج قيمة الدقة 82.50%، والدقة 81.58%، والتذكر 89.39%، والنتيجة f1–نتيجة 85.31% في النموذج A، وقيمة الدقة 81.65%، والدقة 80.73%، والتذكر 89.06%، والنتيجة f1–نتيجة 84.69% في النموذج B.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pola hidup manusia setiap tahun semakin berkembang seiring dengan perkembangan teknologi. Namun, semakin banyak pula jenis penyakit yang berpotensi untuk diderita mereka. Salah satu penyebab muncul berbagai macam penyakit dalam tubuh manusia yaitu pola hidup dan pola makan yang tidak sehat, serta pengaruh asap rokok. Asap rokok tersebut salah satu residu yang dihasilkan oleh perokok. Perilaku merokok dapat memberikan banyak sekali dampak buruk yang dapat terjadi secara langsung maupun dampak yang buruk beresiko jangka panjang. Salah satu dampak tersebut berpengaruh pada kondisi tubuhnya maupun kondisi tubuh orang lain yang bisa berakibat jangka panjang. Dan dampak terbesarnya dapat mengakibatkan munculnya penyakit kronis seperti *stroke*, penyakit jantung, kanker, gangguan pernapasan serta kemungkinan paling parah dapat menyebabkan kematian (Najiyah et al., 2020).

Menurut *World Health Organization* (*WHO*), secara global lebih dari 22000 orang meninggal dunia setiap harinya yang diakibatkan oleh paparan asap rokok. Dalam paparan asap rokok tersebut terkandung banyak sekali zat berbahaya. Nikotin adalah salah satu zat berbahaya yang bersifat racun pada saraf manusia. Selain Nikotin, Tar juga termasuk ke dalam zat berbahaya rokok yang bersifat karsinogenik dan memberikan efek iritasi serta menimbulkan kanker pada organ pernapasan. Zat lainnya yaitu gas karbon monoksida (CO) yang dihasilkan dari

aktivitas merokok juga dapat mempengaruhi kinerja sel-sel darah merah dalam mengikat oksigen (Aji et al., 2015).

Zat-zat berbahaya dalam asap rokok dapat terdeteksi melalui pemeriksaan tubuh. Dalam pemeriksaan kondisi tubuh, umumnya ditemukan kandungan zat karbon monoksida (CO) yang terdapat dalam darah melalui tes darah. Dalam hal ini dapat diketahui seseorang termasuk perokok atau tidak, namun perlu pula untuk melihat aspek berpengaruh lainnya. Dapat pula diketahui dengan kadar hemoglobin yang meningkat dalam darah yang dipengaruhi oleh nikotin yang memicu pembentukan sel darah merah lebih banyak (Ulandhary et al., 2020). Kandungan nikotin dan gas karbon monoksida yang terdapat dalam rokok dapat dipastikan bisa menyebabkan efek yang lebih besar apabila dihirup dalam waktu yang lama. Dengan mengetahui efek rokok pada kondisi tubuh perokok maka penting untuk dilakukan upaya pencegahan dan penanganan, sehingga ahli medis yang melakukan perawatan terhadap pasien, dapat segera memutuskan perawatan yang cocok terhadap pasien yang terpapar asap rokok maupun pasien yang sengaja merokok. Sehingga usaha untuk membantu untuk penanganan pasien dengan tepat, dapat sesuai sebagaimana firman Allah SWT dalam penggalan surah Al-Maidah ayat 2 yang berbunyi:

وَتَعَاوَنُوْا عَلَى الْبِرِّ وَالتَّقْوٰى وَلَا تَعَاوَنُوْا عَلَى الْاِثْمِ وَالْعُدْوَانِ وَاتَّقُوا اللّٰهَ اِنَّ اللّٰهَ شَدِيْدُ الْعِقَابِ

"Dan tolong-menolonglah kamu dalam mengerjakan kebajikan dan takwa, dan jangan tolong menolong dalam mengerjakan perbuatan dosa dan permusuhan. Bertakwalah kepada Allah, sesungguhnya Allah sangat berat siksaan-Nya", (Q.S al Maidah ayat 2)

Dalam potongan ayat tersebut Allah SWT. Memerintahkan kepada hambanya agar saling tolong-menolong dalam hal kebaikan dan tidak saling tolong-menolong dalam hal keburukan. Sehingga dengan pelaksanaan penelitian ini ahli medis dapat dengan baik menjalankan perintah Allah SWT dalam membantu penyembuhan penyakit yang dialami oleh pasien. Penelitian ini ditujukan untuk memerikan kontribusi positif dalam pengembangan pengetahuan medis dan praktik kesehatan, sehingga setiap tindakan medis yang dilakukan dapat sesuai dengan nilai-nilai etika islam. Dengan demikian, pasien dapat mendapatkan perawatan yang tidak hanya didasarkan pada ilmu kesehatan modern, tetapi juga mencerminkan kepatuhan terhadap ajaran agama khususnya dalam hal tolong menolong.

Selama beberapa tahun terakhir, *decision tree* menjadi semakin umum digunakan sebagai metode prediksi dan klasifikasi dalam bidang analisis data. *Decision tree* merupakan model prediktif yang berfungsi untuk memberikan pemahaman intuitif tentang sebuah proses pengambilan keputusan. *Decision tree* dapat menangani kompleksitas data dengan efektif, namun metode ini memiliki kelemahan seperti kecenderungan untuk overfitting. Sebagai alternatif dari masalah tersebut, *Random Forest* muncul sebagai pengembangan dari *Decision tree* untuk mengatasi kelemahannya. Cara kerja dari *Random Forest* adalah dengan menggabungkan prediksi dari beberapa *decision tree* yang dibuat secara acak, meningkatkan akurasi dan mengurangi risiko terjadinya overfitting. Sebagai contoh, hasil dari salah satu penelitian yang melakukan perbandingan antara *Decision tree* dengan *Random Forest* dalam menentukan faktor risiko pada

Diabetes Tipe 2 menunjukkan bahwa nilai akurasi dari model *Random Forest* lebih tinggi dari pada model *Decision tree* (Esmally et al., 2018).

1.2 Rumusan Masalah

Bagaimana *Performance* metode *Random Forest* yang baik melalui perhitungan Accuracy, Precision, Recall, dan f1-score untuk klasifikasi perokok berdasarkan kondisi tubuh?

1.3 Batasan Masalah

1. Penelitian ini menerapkan metode *Random Forest Classifier*
2. Pada penelitian ini hanya menggunakan satu *Dataset* dengan judul *Body Signal of Smoking* yang berasal dari Departemen Kantor Pusat Strategi *Big Data* di Perusahaan Asuransi Kesehatan Nasional Republik Korea.
3. Penelitian ini tidak membahas lebih detail tentang penyakit yang disebabkan merokok.
4. Penelitian ini menggunakan bahasa pemrograman Python
5. Penelitian ini menggunakan Google Colab sebagai *Interactive Notebook Environment* untuk menjalankan kode Python.

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui *Performance* metode *Random Forest* dengan baik dalam mengklasifikasikan perokok berdasarkan kondisi tubuh dengan menerapkan *Random Forest*.

1.5 Manfaat Penelitian

1. Penelitian ini dapat menghasilkan model prediktif yang dapat digunakan untuk memprediksi kecenderungan seseorang sebagai perokok berdasarkan kondisi tubuh mereka. Model ini dapat membantu dalam pengembangan strategi pencegahan dan intervensi yang lebih personal dan efektif.
2. Penelitian ini akan meningkatkan pemahaman tentang hubungan antara kondisi tubuh dengan kebiasaan merokok sehingga dapat memberikan wawasan baru dan dapat membantu ahli medis menemukan perawatan yang tepat untuk setiap individu.

BAB II

STUDI PUSTAKA

2.1 Penelitian Terkait

Dalam penelitian (Hidayat et al., 2023) tentang Klasifikasi penyakit jantung menggunakan *Random Forest clasifier*, menggunakan pengujian sebanyak dua belas kali dengan hasil terbaik pada pengujian pertama yang ke-6 dengan menggunakan pembagian data 80%:20% yang menghasilkan akurasi sebanyak 94%. Hasil tersebut lebih baik dibandingkan ketika data dibagi menjadi 70%:30% yang menghasilkan akurasi sebesar 92%. Berdasarkan penelitian tersebut, pembaruan yang akan dilakukan pada penelitian ini yaitu dengan menggunakan *Dataset Body Signal of Smoking*, serta penggunaan *tuning Hyperparameter* menggunakan *GridSearchCV*.

Penelitian tentang Klasifikasi kualitas air sumur menggunakan algoritma *Random Forest* yang dilakukan oleh Muhamad Malik Mutoffar (Mutoffar & Fadillah, 2022) menghasilkan nilai akurasi sebesar 82% dengan pembagian data sebesar 80% pada data *train* dan 20% untuk data *test*. Dalam data tersebut dapat dilakukan klasifikasi pada air yang dapat dikonsumsi dan air yang tidak dapat dikonsumsi. Hasil klasifikasi tersebut berasal dari 267 data yang telah dikumpulkan dan terdapat 40 data yang berhasil diprediksi nilainya sesuai dengan targetnya. Berdasarkan penelitian tersebut, pembaruan yang akan dilakukan pada penelitian ini yaitu menambahkan satu model perbandingan data *train* dan data *test* dengan rasio 70%:30%.

Dalam penelitian yang dilakukan oleh (Azizah et al., 2023) yang berjudul Model Klasifikasi Berbasis Ekspresi Gen *Non-Small Cell Lung Carcinoma* (NSCLC) pada Wanita Bukan Perokok Menggunakan Metode *Ensemble*. Tujuan dari penelitian tersebut bermaksud untuk memprediksi NSCLC menggunakan metode *ensemble* pada data *microarray*. Metode *ensemble* yang digunakan yaitu *Random Forest*, *Adaptive Boosting*, dan *Extreme Gradient Boosting* yang digunakan untuk memprediksi NSCLC. Berdasarkan penelitian tersebut, pembaruan yang akan dilakukan pada penelitian ini yaitu dengan menambahkan *hyperparameter* yang digunakan diantaranya *n_estimators*, *max_depth*, *bootstrap* yang bernilai *false*, dan *random_state*.

Berdasarkan penelitian yang dilakukan oleh (Issabakhsh et al., 2023) yang berjudul *Machine Learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study*, yang melakukan penelitian untuk mengidentifikasi faktor penentu penghentian merokok diantara perokok dewasa yang ada di AS. Dalam prosesnya diperoleh hasil akhir prediksi penghentian merokok dengan akurasi 71% pada gelombang 1 dan akurasi 70% pada gelombang 2. Model *Machine Learning* yang digunakan adalah *Random Forest*. Berdasarkan penelitian tersebut, pembaruan yang akan dilakukan pada penelitian ini yaitu dengan menambahkan skenario pengujian berupa metode penyeimbangan data menggunakan *SMOTE* dan Seleksi Fitur menggunakan *Feature importances*.

Menurut (Song et al., 2021) dalam penelitiannya yang berjudul "*The Random Forest Model Has the Best Accuracy Among the Four Pressure Ulcer Prediction Models Using Machine Learning Algorithms*" mengevaluasi empat

model *Machine Learning* yaitu *Support Vector Machine, Decision tree, Random Forest* dan ANN yang digunakan untuk memprediksi efek samping ulkus tekanan dan menemukan kemampuan tertinggi masing-masing model dalam memprediksi efek samping ulkus tekanan. Dalam prosesnya, ditemukan bahwa *Random Forest* dan *Decision tree* memiliki kinerja prediksi yang lebih baik dan lebih cocok untuk masalah tersebut. Perbedaan dalam hal *Dataset* yang digunakan merupakan data primer yang secara langsung diambil di sebuah rumah sakit. Sedangkan *Dataset* yang digunakan oleh peneliti menggunakan data yang telah tersedia pada *kaggle* dengan judul *Body Signal of Smoking*. Pembaruan lain yang dilakukan pada penelitian ini yaitu implementasi *Synthetic Minority Oversampling Technique* (*SMOTE*) dan Seleksi Fitur menggunakan *Feature importances* serta penambahan beberapa *hyperparameter* seperti *n_estimators*, *max_depth*, *bootstrap* yang bernilai *false,* dan *random_state*.

Dalam penelitian (Adian et al., 2023) tentang Implementasi *Random Forest* pada klasifikasi penyakit Kardiovaskular dengan *Hyperparameter Tuning GridSearchCV*, dilakukan klasifikasi menggunakan metode *Random Forest* yang mengimplementasikan model pengujian tanpa *tuning Hyperparameter* dan model pengujian dengan t*uning Hyperparameter GridSearchCV*. Penelitian tersebut menghasilkan nilai *Performance* yang lebih baik dengan menggunakan *GridSearchCV*. Pembaruan pada penelitian yang akan dilakukan yakni menggunakan *Dataset Body Signal of Smoking*, serta menggunakan *Hyperparameter* yang berbeda serta penambahan skenario pengujian *Synthetic*

Minority Oversampling Technique (*SMOTE*) untuk penyeimbangan data dan Seleksi Fitur menggunakan *Feature importances*.

Tabel 2. 1 Perbandingan Penelitian

No	Referensi	Metode Penelitian	Hasil Penelitian	Perbedaan
1	(Hidayat et al., 2023)	*Random Forest*	Menghasilkan akurasi terbaik sebesar 94% menggunakan perbandingan data *train* dan data *test* sebanyak 80%:20%. Hasil akhir tersebut lebih besar nilainya dibandingkan dengan ketika menggunakan perbandingan data *train* dan data *test* sebanyak *7*0%:30% yang memberikan hasil sebesar 92%.	Data yang digunakan dalam penelitian tersebut yaitu *Dataset* penyakit jantung. Sedangkan peneliti menggunakan *Dataset Body Signal of Smoking*. Perbedaan pada penggunaan *tuning Hyperparameter*, penelitian tersebut menggunakan tuning hyperparameter secara manual, sedangkan pada penelitian ini akan menggunakan *GridSearchCV*.
2	(Mutoffar & Fadillah, 2022)	*Random Forest*	Penelitian ini menghasilkan nilai akurasi sebesar 82% dengan perbandingan data *train* dan data *test* sebanyak 80%:20%.	Perbedaan pada objek yang diteliti serta dataset yang akan digunakan pada penelitian mendatang akan diuji dengan menambahkan satu model perbandingan data yaitu perbandingan data *train* dan data *test* 70%:30%.
3	(Azizah et al., 2023)	*Random Forest, adaptive boosting, extreme gradient boosting*	Dalam prosesnya yang menggunakan tiga algoritma yang berbeda dengan fitur yang digunakan berbeda pula. Hasil dari ketiganya adalah *Random Forest 0.93, adaptive boosting 1.00*, dan *extreme gradient boosting 0.93*.	Perbedaan dataset yang digunakan, serta terdapat penambahan beberapa *hyperparameter Random Forest* yang berbeda dari penelitian sebelumnya.
4	(Issabakhsh et al., 2023)	*Random Forest, Gradient Boosting Machine*	Hasil penerapan kedua algoritma tersebut memberikan nilai akurasi sebesar 71%.	Perbedaan objek penelitian dan penambahan skenario pengujian dengan *SMOTE* dan seleksi

No	Referensi	Metode Penelitian	Hasil Penelitian	Perbedaan
				fitur menggunakan *Feature importances.*
5	(Song et al., 2021)	*Support vector machine, Decision tree, Random Forest, ANN*	Dalam prosesnya, ditemukan bahwa *Random Forest* dan *decision tree* memiliki kinerja prediksi yang lebih baik dan lebih cocok untuk masalah prediksi efek samping ulkus tekanan.	Perbedaan dalam hal *Dataset* yang digunakan merupakan data primer yang secara langsung diambil di sebuah rumah sakit. Sedangkan *Dataset* yang digunakan oleh peneliti menggunakan data yang telah tersedia pada kaggle dengan judul *Body Signal of Smoking.*
6	(Adian et al., 2023)	*Random Forest*	Dalam penelitian tersebut diterapkan model tanpa *tuning Hyperparameter* dan menggunakan *tuning Hyperparameter* dengan *GridSearchCV,* dan menghasilkan *Performance* yang lebih baik saat menggunakan *tuning Hyperparameter.*	Perbedaan pada objek penelitian (*Dataset Body Signal of Smoking*), menggunakan *Hyperparameter* yang berbeda serta penambahan skenario pengujian *SMOTE* dan Seleksi fitur menggunakan *Feature importances*.

2.2 *Dataset*

Menurut IBM, *Dataset* dapat diartikan sebagai kumpulan data. Istilah tersebut mengacu pada file yang berisi satu atau lebih data. Sedangkan data merupakan sebuah informasi yang bisa mendeskripsikan berbagai macam hal, diantaranya seperti deskripsi, statistik, dan banyak jenis data yang lain. Umumnya, *dataset* hanya mencakup satu topik saja. Fungsi dari *dataset* adalah sebagai catatan yang dapat dimanipulasi dan diolah menjadi bentuk *dataset* baru (Syafrina, 2018). Berdasarkan jenisnya, *dataset* dibedakan menjadi dua, yaitu *private dataset* dan *public dataset*. *Private dataset* ialah *dataset* yang bersumber dari organisasi tempat peneliti melakukan penelitian, sehingga hanya orang tertentu yang memiliki

aksesnya. *Public dataset* merupakan *dataset* yang dapat diambil dan bersumber dari *repository public* yang disepakati oleh para peneliti(Yazid et al., 2022).

2.3 Perokok

Merokok merupakan salah satu perilaku yang merugikan bagi orang yang melakukannya maupun orang lain yang menghirup asap yang dihasilkan. Merokok dapat menyebabkan denyut jantung bekerja lebih cepat untuk mengirimkan oksigen. Hal tersebut juga dapat menurunkan level HDL kolesterol dalam darah. HDL atau *High Density Lipoprotein* adalah kolesterol baik yang bertugas mengangkut kolesterol jahat berlebih yang terdapat pada dinding arteri ke hati yang kemudian akan dikeluarkan melalui saluran pencernaan. Sebagian besar perokok memiliki kadar HDL lebih rendah daripada orang normal yang tidak merokok maupun perokok pasif (Najiyah et al., 2020). Rendahnya kadar HDL dapat menimbulkan penyakit jantung koroner yang disebabkan kelebihan kolesterol yang menyumbat pembuluh darah akibat tidak terangkut sempurna dan dapat menghambat pasokan oksigen ke jantung karena meningkatnya tekanan dalam pembuluh darah (Kholidha et al., 2019).

Faktor risiko yang diciptakan aktivitas merokok sangat signifikan untuk morbiditas dan mortalitas serta merupakan penyebab paling penting yang dapat dicegah dari berbagai penyakit. Merokok tidak dapat menyebabkan disfungsi endotel atau lapisan sel yang melapisi bagian dalam pembuluh darah. Namun, merokok dapat dikaitkan dengan *dislipidemia* atau ketidakseimbangan kadar *lipid* (lemak) dalam darah. Hal tersebut ditandai dengan peningkatan trigliserida yang berfungsi menyediakan energi bagi tubuh dan mendukung kinerja tubuh yang lain.

Meingkatnya kadar trigliserida tersebut dapat menyebabkan terjadinya penyakit kardiovaskular. Merokok juga dapat meningkatkan kadar LDL (*Low-Density Lipoprotein*) atau kolesterol jahat dalam tubuh sehingga dapat meningkatkan risiko terpapar penyakit kardiovaskular. Sebagaimana umumnya, LDL berperan sebagai transportasi kolesterol dan lemak dalam darah, apabila jumlahnya diluar batas maka dapat berpengaruh pada keseimbangan kinerja tubuh(Nakamura et al., 2021). Berdasarkan penelitian yang dilakukan oleh (Malaeny et al., 2017) dijelaskan tentang kondisi perokok yang memiliki kadar kolesterol total yang lebih tinggi daripada non-perokok.

Menurut Permatasari dalam (Arifin & Yunasri, 2021) menyebutkan bahwa karbon monoksida yang terhirup dan masuk ke dalam tubuh dapat menghambat kinerja hemoglobin untuk mengikat oksigen. Hemoglobin adalah komponen dalam darah yang berfungsi untuk mengangkut oksigen dari paru-paru ke seluruh organ jaringan pada tubuh. Kadar hemoglobin yang rendah dalam darah dapat mengakibatkan suplai oksigen pada organ-organ tubuh menjadi berkurang, yang dapat menyebabkan seseorang mengalami anemia(Muzayyaroh & Suyati, 2018).

Menurut *National Energy and Climate Plans* (NECP) dalam penelitian yang dilakukan oleh (Ekayanti, 2019) mengatakan kadar kolesterol total yang normal adalah kurang dari 200 mg/dL, sedangkan apabila lebih dari 200 mg/dL perlu diwaspadai agar tidak mengakibatkan masalah kesehatan yang lebih serius. Dalam penelitian (Kusumasari, 2015) ditemukan rata-rata kadar kolesterol total milik perokok mencapai 250,5 mg/dL, sedangkan kadar kolesterol total non perokok rata-rata 166,2 mg/dL.

Serum AST atau *Asparat Aminotranferase* merupakan enzim yang ditemui pada otot jantung dan hati (Nasution, 2022). Normalnya enzim ini sebanyak 10-45 U/L dalam tubuh. Serum ALT atau *Alanin Aminotranferase* adalah enzim yang mayoritas ditemukan pada sel hati dan efektif dalam mendiagnosis destruksi hepatoseluler. Kadar normal enzim ALT untuk tubuh sekitar 5-45 U/L (Khatri et al., 2021).

2.4 *Machine Learning*

Beberapa tahun yang lalu, *Machine Learning* menjadi salah satu cabang dari artificial intelligence yang sangat berperan penting dalam dunia IT (Bonaccorso, 2017). *Machine Learning* merupakan teknik yang berperan meningkatkan kinerja sistem dengan belajar dari pengalaman melalui metode komputasi. Dalam sistem komputer, pengalaman ini direpresentasikan dalam bentuk data, dan tujuan utama *Machine Learning* adalah mengembangkan algoritma pembelajaran yang membangun model dari data tersebut. Dengan memberikan data pengalaman ke algoritma pembelajaran, kita dapat menciptakan model yang mampu membuat prediksi pada data baru yang belum pernah dilihat sebelumnya (Zhou, 2021).

Dalam *Machine Learning* terdapat teknik ensemble learning yang memberikan alternatif kuat dalam algoritma kompleks dengan mencoba mengeksploitasi konsep statistik mayoritas suara. Metode ensemble bukanlah satu-satunya metode yang paling umum digunakan dalam klasifikasi. Namun, metode ini menawarkan tingkat kesederhanaan yang baik dan dapat digunakan dalam banyak tugas yang tidak membutuhkan tingkat kerumitan yang tinggi. Metode tersebut sangat berguna ketika diperlukan untuk menunjukkan bagaimana sebuah

proses keputusan bekerja. Cara kerja *ensemble learning* adalah dengan menggabungkan beberapa model *Machine Learning* yang diperlukan untuk membuat prediksi dan keputusan yang lebih akurat dan stabil dibandingkan dengan yang dihasilkan setiap model secara individu (Bonaccorso, 2017).

Penerapan dari algoritma *machine learning* yang akan digunakan untuk penelitian ini adalah algoritma klasifikasi. Klasifikasi ini berfungsi untuk melakukan pengelompokan atau mengkategorikan data dalam kelas atau kategori yang telah ditentukan berdasarkan karakteristik atau atribut tertentu. Algoritma ini umumnya membangun model yang dapat memprediksi kelas dari data yang belum dilihat sebelumnya menggunakan pembelajaran dari data yang sudah dilabeli atau diketahui kelasnya dan termasuk dalam *supervised learning* (Rahmadeyan & Mustakim, 2023). Salah satu algoritma yang akan digunakan adalah algoritma *Random Forest*.

2.5 *Exploratory Data Analysis* (*EDA*)

Exploratory Data Analysis (*EDA*) merupakan sebuah pendekatan untuk menganalisis data yang berfokus pada pemahaman karakeristik utama data dan mengidentifikasi pola dan hubungan di dalamnya. Beberapa komponen yang terdapat dalam *EDA* adalah eksplorasi data, visualisasi data, prapemrosesan data, analisis statistik, dan pengenalan pola (Yu, 2015).

1. Eksplorasi data

   Dalam proses ini melibatkan pemeriksaan data untuk memahami struktur, distribusi, dan hubungan antar variabel.

2. Visualisasi data

   EDA sering mencakup dalam pembuatan representasi visual seperti plot, bagan, dan grafik untuk mengidentifikasi pola, tren, dan *outlier* dalam data.

3. Preprocessing data

   Dalam fitur ini melibatkan pembersihan dan transformasi data, penanganan nilai yang hilang, *outlier,* dan mengatasi inkonsistensi data.

4. Analisis statistik

   Fitur ini menggunakan berbagai jenis teknik statistik untuk meringkas dan menggambarkan data, termasuk ukuran kecenderungan sentral, dispersi, korelasi, dan pengujian hipotesis.

5. Pengenalan pola

   Teknik yang digunakan dalam fitur ini seperti jaringan saraf dan pohon klasifikasi untuk mengidentifikasi dan mengekstrak pola atau hubungan yang bermakna dalam data.

2.6 *Synthetic Minority Oversampling Technique* (*SMOTE*)

Dalam permasalahan data yang tidak seimbang, maka dilakukan proses penyeimbangan data. Teknik penyeimbangan data dilakukan dengan dua cara yaitu:

a. Undersampling

Teknik ini menyeimbangkan dataset dengan mengurangi ukuran sampel dalam kelas mayoritas agar seimbang dengan kelas minoritas. Pendekatan ini menyebabkan kehilangan informasi yang signifikan pada data dari kelas mayoritas, namun dapat efektif dalam beberapa kasus kelas mayoritas yang memiliki banyak sampel.

b. Oversampling

Teknik *oversampling* bekerja dengan cara meningkatkan jumlah sampel dalam kelas minoritas agar seimbang dengan kelas mayoritas. Cara yang digunakan adalah mengulangi sampel-sampel yang ada dalam kelas mayoritas menggunakan *duplicating instances*, atau *SMOTE* yang menciptakan sampel sintetis baru dalam kelas minoritas untuk menyeimbangkan proporsi kelas dalam dataset.

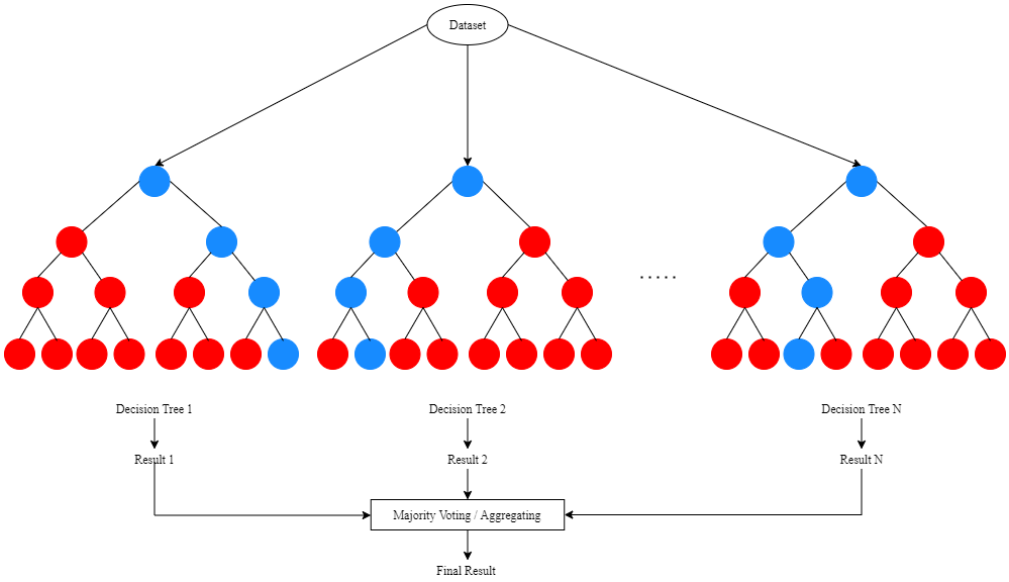
Penelitian ini menggunakan *SMOTE* untuk mengatasi masalah *imbalance data*. Salah satu keuntungan *SMOTE* adalah mengurangi risiko terjadinya *overfitting* yang sering terjadi ketika menggunakan pendekatan *oversampling* sederhana yang hanya mengulangi sampel-sampel minoritas yang sudah ada.

2.7 *Random Forest*

Sebelum melakukan implementasi algoritma *Random Forest*, terlebih dahulu dilakukan *split data* pada *dataset* yang bertujuan sebagai pelatihan dan pengujian. Manfaatnya yakni dapat memberikan perkiraan model yang lebih kredibel karena diuji pada data yang tidak terlihat sebelumnya. Dalam tahapan ini, *dataset* akan dibagi menjadi data *train* dan data *test. Dataset* akan dilakukan pengujian menggunakan dua model yang berbeda berdasarkan perbandingan data *train* dan data *test*-nya. Dalam penelitian ini dilakukan dengan dua model berbeda, sehingga hasil prediksi yang akan didapatkan dapat menghasilkan nilai ketepatan yang terbaik (Teguh et al., 2022).

Metode *Random Forest* merupakan pengembangan dari algoritma *Decision tree*. Dalam prosesnya terdapat perbedaan pada pengambilan sampel secara acak,

sesuai dengan istilahnya *Random* yang berarti “acak”. Sehingga terdapat banyak *Decision tree* yang menghasilkan beragam hasil. Masing-masing *decision tree* tersebut dibuat menggunakan pemilihan atribut secara acak pada tiap simpulnya untuk menentukan pembagiannya. Secara lebih formal, setiap *tree* akan bergantung pada nilai vektor acak yang disampel secara independen dan dengan distribusi yang sama untuk semua pohon dalam *forest* tersebut. Selama proses klasifikasi, setiap pohon akan memberikan keputusan yang berbeda dan keputusan mayoritas yang akan dipilih (Han et al., 2022). Secara lebih detail, *Random Forest* akan bekerja seperti ilustrasi yang terdapat dalam gambar 2.1 berikut.



Gambar 2. 1 Ilustrasi alur kerja *Random Forest*

Random Forest termasuk dalam salah satu algoritma *Machine Learning* yang menerapkan teknik *ensemble*. Teknik *ensemble* adalah sebuah teknik yang menggabungkan beberapa model untuk membuat prediksi dengan lebih akurat daripada satu model saja. Pada prosesnya *Random Forest* menggunakan *Bagging* atau *Bootstrap Aggregating* yang berfungsi untuk meningkatkan kinerja dan

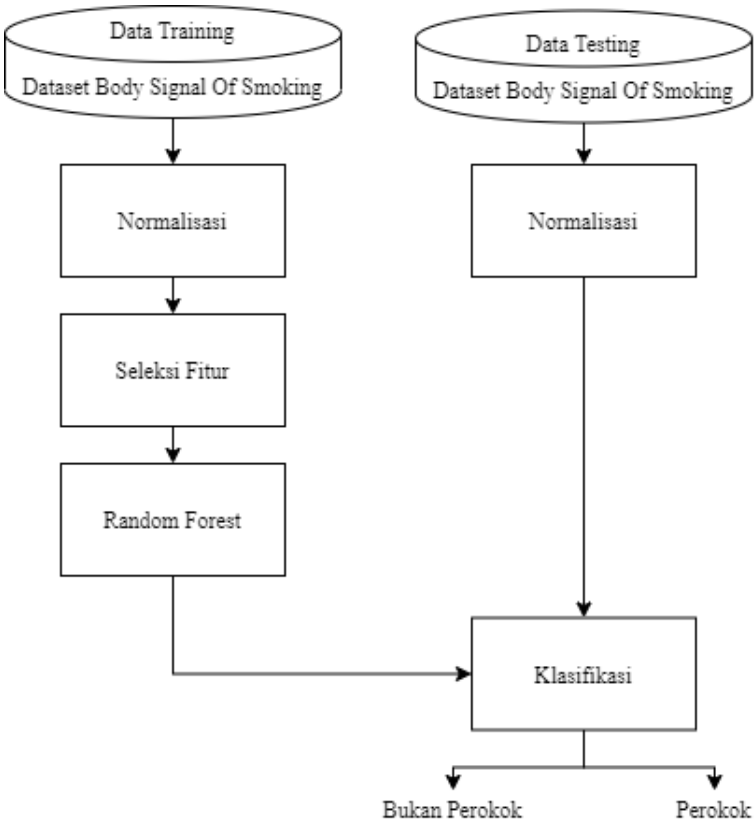
kestabilan model *Machine Learning*. Metode ini menerapkan pemilihan fitur secara acak disetiap pohonnya sehingga dapat meningkatkan akurasinya dalam melakukan klasifikasi dan regresi. Jenis keacakan yang tepat memastikan bahwa masing-masing pohon tidak terlalu bergantung satu sama lain, mengurangi korelasi dan meningkatkan kinerja keseluruhan pola *Random Forest* (Breiman, 2001). Sebuah model klasifikasi dapat dikategorikan sangat baik saat model tersebut mencapai nilai accuracy sebesar 90%-100%, model klasifikasi tergolong baik saat 80%-90%, kategori klasifikasi cukup baik pada nilai 70%-80%, 60%-70% termasuk klasifikasi kurang baik dan klasifikasi gagal saat nilai dalam rentang 50%-60% (Subarkah et al., 2022).

BAB III

METODOLOGI PENELITIAN

3.1 Desain Sistem

Tahap desain sistem merupakan tahap yang digunakan peneliti untuk merencanakan alur penelitian yang terstruktur. Alur penelitian tersebut dapat dilihat dalam gambar 3.1 dibawah ini.



Gambar 3. 1 Desain Sistem

3.2 Pengumpulan Data

Data yang digunakan dalam penelitian ini menggunakan data sekunder yang diambil dari *website* kaggle dengan judul *Body Signal of Smoking*. *Dataset* tersebut memiliki 26 atribut didalamnya dan satu target yaitu *Smoking* yang memiliki *Class*

yaitu 0 (*non-smoking*) dan 1 (*smoking*). Data ini berasal dari data pemerintahan negara korea selatan yang berisi tentang data kesehatan sebagian penduduknya. Secara lebih detail, data tersebut diperoleh dari salah satu Perusahaan Asuransi Kesehatan Nasional milik Republik Korea Selatan pada tahun 2022. Dalam dataset tersebut berisi 55692 baris data yang berasal dari hasil pemeriksaan penduduk yang menjadi penerima asuransi kesehatan. Berikut beberapa atribut yang terdapat didalamnya:

Tabel 3. 1 Detail Atribut dalam Dataset

No	Atribut	Keterangan
1	*ID*	ID yang terdapat dalam
2	*Gender*	Jenis kelamin
3	*Age*	Usia pasien dalam tahun
4	*Height*	Tinggi badan dalam satuan centimeter (cm)
5	*Weight*	Berat badan dalam satuan kilogram (kg)
6	*Waist (cm)*	Panjang lingkar pinggang
7	*Eyesight (left)*	Pengukuran penglihatan pada mata kiri
8	*Eyesight (right)*	Pengukuran penglihatan pada mata kanan
9	*Hearing (left)*	Pengururan pendengaran pada telinga kiri
10	*Hearing (right)*	Pengukuran pendengaran pada telinga kanan
11	*Systolic*	Tekanan dalam arteri ketika jantung memompa darah
12	*Relaxation*	Proses pemulihan otot jantung ke kondisi awal setelah kontraksi.
13	*Fasting blood sugar*	Kadar gula darah setelah tidak makan selama semalam
14	*Cholesterol*	Total kolesterol
15	*Triglyceride*	Level trigliserida dalam tubuh
16	HDL	Dapat disebut kolesterol baik
17	LDL	Kolesterol jahat
18	Hemoglobin	Kadar hemoglobin dalam tubuh
19	*Urine protein*	Kadar protein dalam urin
20	*Serum creatinine*	Kadar kreatinin dalam darah
21	AST	*Aspartat aminotransferase*, enzim yang banyak ditemukan di hati
22	ALT	*Alanin transaminase*, enzim yang sebagian besar ditemukan di hati
23	Gtp	Nukleotida kaya energi yang analog dengan ATP
24	Oral	Status pemeriksaan mulut
25	*Dental caries*	Ada tidaknya karies gigi
26	Tartar	Status karang gigi

Tabel 3. 2 Sampel Dataset

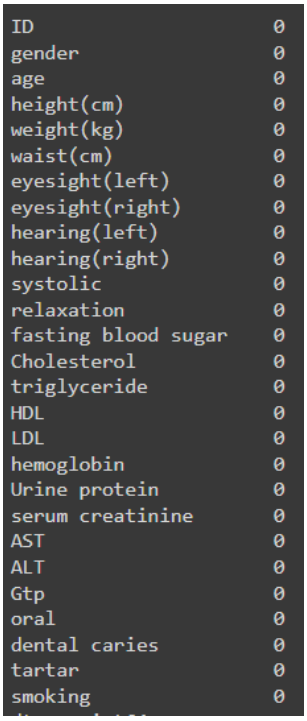
ID	***gender***	***age***	***height (cm)***	***weight (kg)***	***waist (cm)***	***eyesight (left)***	***eyesight (right)***	***hearing (left)***	***hearing (right)***	***systolic***	***relaxation***	***Fasting blood sugar***	***cholesterol***
0	0	40	155	60	81,3	1,2	1	1	1	114	73	94	215
1	0	40	160	60	81	0,8	0,6	1	1	119	70	130	192
2	1	55	170	60	80	0,8	0,8	1	1	138	86	89	242
3	1	40	165	70	88	1,5	1,5	1	1	100	60	96	322
4	0	40	155	60	86	1	1	1	1	120	74	80	184

triglyceride	***HDL***	***LDL***	***hemoglobin***	***Urine protein***	***Serum creatinine***	***AST***	***ALT***	***gtp***	***oral***	***Dental caries***	***tartar***	***smoking***
82	73	126	12,9	1	0,7	18	19	27	0	0	1	0
115	42	127	12,7	1	0,6	22	19	18	0	0	1	0
182	55	151	15,8	1	1	21	16	22	0	0	0	1
254	45	226	14,7	1	1	19	26	18	0	0	1	0
74	62	107	12,5	1	0,6	16	14	22	0	0	0	0

3.3 *Exploratory Data Analysis* (*EDA*)

3.3.1 Mencari *missing value*

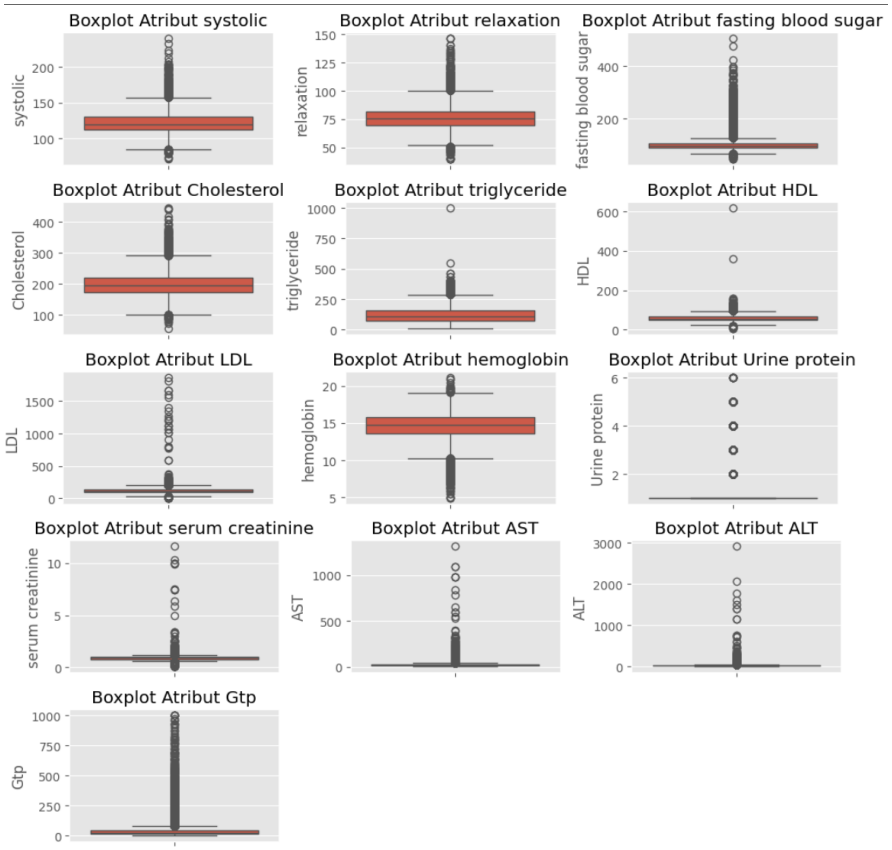
Dalam proses ini dilakukan pencarian nilai null pada semua baris data. Dengan menggunakan *library Pandas*, fungsi ‘isnull’ dan ‘sum()’ dapat membantu dalam identifikasi dan analisis data yang hilang atau *missing value*. Hasil dari pencarian tersebut menunjukkan bahwa *dataset* yang digunakan tidak memiliki *missing value* yang ditunjukkan pada gambar 3.2



Gambar 3. 2 Pencarian *Missing Values*

3.3.2 Penghapusan *outliers*

Salah satu hal yang perlu dilakukan untuk preprocessing data adalah penghapusan *outlier*. Sebelum penghapusan data terlebih dahulu dilakukan proses pemeriksaan *outlier* dengan *Boxplot*. Gambar 3.3 sebagai salah satu atribut yang memiliki *outlier*.



Gambar 3. 3 Atribut yang memiliki *outlier*

Salah satu metode yang digunakan untuk menghapus *outlier* dapat menggunakan *Interquartile Range* (IQR) yang didefinisikan pada persamaan 3.1 berikut:

$$IQR = Q3 - Q1 \tag{3.1}$$

IQR = *Interquartile Range*
$Q3$ = Quartil atas
$Q1$ = Quartil bawah

Dengan Q3 sebagai nilai tengah antara median dan nilai maksimum dari data yang telah diurutkan dan Q1 adalah nilai tengah antara nilai terkecil atau minimum

dan median dari data yang telah diurutkan. Untuk mengidentifikasi batas bawah dan batas atas untuk deteksi *outlier* ditentukan dengan persamaan 3.2 dan 3.3 berikut:

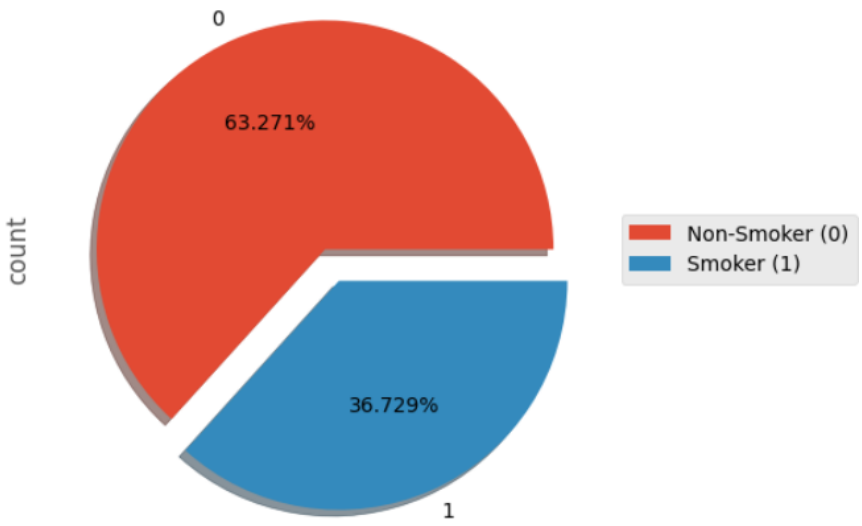
$$Batas\ Bawah = Q1 - (1{,}5 \times IQR) \quad (3.2)$$

$$Batas\ Atas = Q3 + (1{,}5 \times IQR) \quad (3.3)$$

Sehingga nilai yang berada diluar batas tersebut sering dianggap sebagai *outliers*. *Outlier* yang ditemukan dalam dataset penelitian ini sebanyak 616 data.

3.3.3 Mengatasi *Imbalance Data* menggunakan *SMOTE*

Nilai yang terdapat dalam dataset tidak selalu memberikan nilai yang seimbang. Beberapa memiliki nilai target yang lebih banyak pada salah satu kategori. *Dataset* yang memiliki label target yang tidak seimbang dapat mempengaruhi hasil akurasi model yang dibentuk. Dalam gambar 3.3 dapat dilihat distribusi target *dataset* yang digunakan dalam penelitian ini.



Gambar 3. 4 Distribusi target

Data yang tidak seimbang tersebut, kemudian akan dilakukan proses penyeimbangan data menggunakan *Synthetic Minority Oversampling Technique* (*SMOTE*). Cara kerja *SMOTE* untuk mengurangi risiko terjadinya *overfitting* yang sering terjadi ketika menggunakan pendekatan *oversampling*. Sehingga label target yang memiliki jumlah minoritas akan dibuat sampel dengan mengulangi sampel-sampel minoritas yang sudah ada.

3.4 *Preprocessing Data*

3.4.1 *Encoding categorical data*

Metode yang digunakan untuk *encoding* data dalam penelitian ini adalah *Label encoding* yang bekerja dengan mengubah setiap kategori menjadi nilai numerik sesuai urutan datanya. Atribut yang melalui proses *encoding categorical data* diantaranya adalah atribut *gender, tartar*, dan *oral*.

3.4.2 Seleksi fitur

Dalam penerapannya, digunakan teknik *Features importance*. Cara kerja *Features importance* yaitu dengan mengidentifikasi dan memilih atribut-atribut yang paling penting dalam dataset. Jumlah atribut yang digunakan setelah melakukan seleksi fitur sebanyak 13 atribut. Atribut-atribut terpenting tersebut ditunjukkan pada tabel 3.4

Tabel 3. 3 Atribut yang terpenting setelah seleksi fitur

Rank	**Fitur**
1	*Gender*
2	*Hemoglobin*
3	*Gtp*
4	*Height(cm)*
5	*Triglyceride*
6	*Waist(cm)*
7	*LDL*

Rank	**Fitur**
8	*ALT*
9	*Cholesterol*
10	*HDL*
11	*Fasting blood sugar*
12	*Systolic*
13	*Age*

3.4.3 *Split* Data

Model A dengan perbandingan data *train* sebanyak 80% dan data *test* sebanyak 20%. Model B perbandingan data *train* sebanyak 70% dan data *test* sebanyak 30% (Hidayat et al., 2023).

Tabel 3. 4 Detail jumlah data yang digunakan

Model	**Jumlah data yang digunakan**			
	Sebelum Preprocesing	**Setelah preprocesing**	**Data *Train***	**Data *Test***
A (80%:20%)	55692	55.076	44.060	11016
B (70%:30%)	55692	55.076	38.553	16523

3.4.4 Normalisasi

Pada penelitian ini normalisasi diterapkan dengan salah satu teknik yaitu *Z-score normalization*, sebagaimana perhitungannya ditunjukkan pada persamaan (3.4). Implementasi *Z-score normalization* pada algoritma klasifikasi menggunakan *Standard Scaler*.

$$X_{norm} = \frac{x-\mu}{\sigma} \quad (3.4)$$

X_{norm} = nilai fitur yang telah dinormalisasi
x = nilai asli dari fitur
μ = rata-rata dari semua nilai dalam fitur
σ = standar deviasi dari semua nilai dalam fitur

Tabel 3. 5 Sampel Dataset setelah Normalisasi dengan *Z-score*

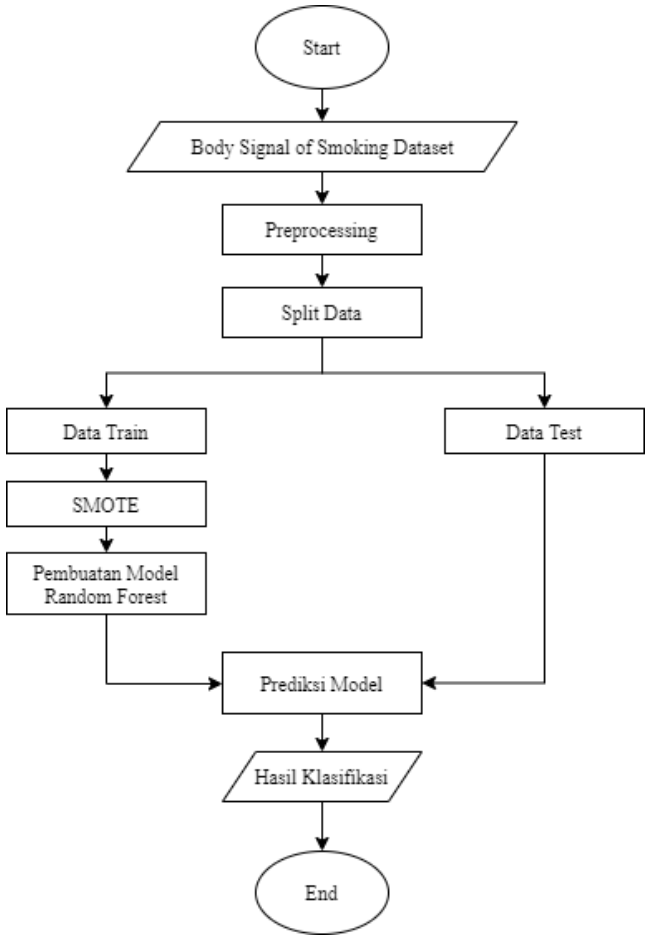
ID	***gender***	***age***	***height (cm)***	***weight (kg)***	***waist (cm)***	***eyesight (left)***	***eyesight (right)***	***hearing (left)***	***hearing (right)***	***systolic***	***relaxation***
0	-1.314542	-0.344847	-1.048314	-0.452484	-0.068258	0.383527	-0.015568	-0.15597	-0.157853	-0.543122	-0.302417
1	-1.314542	-0.344847	-0.504078	-0.452484	-0.100982	-0.439530	-0.858429	-0.15597	-0.157853	-0.172095	-0.616438
2	0.760721	0.900991	0.584395	-0.452484	-0.210062	-0.439530	-0.436998	-0.15597	-0.157853	1.237807	1.058337
3	0.760721	-0.344847	0.040159	0.336805	0,662577	1.000820	1.038009	-0.15597	-0.157853	-1.581998	-1.663172
4	-1.314542	-0.344847	-1.048314	-0.452484	0.444417	-0.028001	-0.015568	-0.15597	-0.157853	-0.097890	-0.197744

Fasting blood sugar	***cholesterol***	***triglyceride***	***HDL***	***LDL***	***hemoglobin***	***Urine protein***	***Serum creatinine***	***AST***
-0.246415	0.508156	-0.618336	1.066699	0.276363	-1.099407	-0.209287	-0.837749	-0.437275
1.536245	-0.130890	-0.151216	-1.042603	0.301235	-1.227721	-0.209287	-1.290384	-0.211827
-0.494006	1.258339	0.797178	-0.158057	0.898166	0.761150	-0.209287	0.520158	-0.268189
-0.147378	3.481105	1.816348	-0.838477	2.763575	0.055422	-0.209287	0.520158	-0.380913
-0.939671	-0.353166	-0.731577	0.318237	-0.196207	-1.356035	-0.209287	-1.290384	-0.549998

ALT	***gtp***	***oral***	***Dental caries***	***tartar***
-0.263860	-0.252364	0.0	-0.515666	0.895078
-0.263860	-0.448376	0.0	-0.515666	0.895078
-0.370689	-0.361260	0.0	-0.515666	-1.117221
-0.014593	-0.448376	0.0	-0.515666	0.895078
-0.441908	-0.361260	0.0	-0.515666	-1.117221

3.5 Implementasi Algoritma *Random Forest*

Dalam tahap implementasi algoritma *Random Forest* untuk klasifikasi perokok, berikut merupakan *flowchart* dari implementasi tersebut sebagaimana dalam gambar 3.5



Gambar 3. 5 *Flowchart Random Forest*

Pada tahap implementasi, masing-masing model akan diterapkan *Hyperparameter*. *Hyperparameter* ialah parameter eksternal yang diperlukan untuk melakukan konfigurasi pada model *Machine Learning* sebelum dilakukan pelatihan. *Hyperparameter* sangat penting untuk dilakukan karena sangat

mempengaruhi performa *Random Forest* (Muhamad & Matin, 2023). *Hyperparameter* yang digunakan dalam pengujian ini terdapat pada tabel 3.6.

Tabel 3. 6 Detail *Hyperparameter*

Hyperparameter	**Nilai**	
n_estimators	100	200
max_depth	12	20
bootstrap	*False*	*False*
random_state	42	42

Hyperparameter n_estimators dan *max_depth* yang terdapat pada tabel 3.4, selanjutnya akan dilakukan pencarian parameter terbaik menggunakan *GridSearch*. Sehingga diperoleh *n_estimators* terbaik dengan nilai 200 dan *max_depth* terbaik dengan nilai 20.

3.6 Evaluasi

Dalam proses evaluasi *Random Forest Classifier*, setelah model *Random Forest* dibangun selanjutnya akan digunakan *Confusion Matrix* untuk tahap evaluasinya. *Confusion Matrix* melibatkan analisis hasil prediksi dari model klasifikasi. *Confusion Matrix* menyajikan informasi tentang klasifikasi yang benar dan salah pada setiap kelas. Komponen utama yang terdapat dalam *Confusion Matrix* adalah *True Positives* (TP), *True Negatives* (TN), *False Positives* (FP) dan *False Negatives* (FN).

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.5}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.7}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.8)$$

3.7 Skenario Pengujian

Dalam penelitian ini dilakukan pengujian dengan dua model sebagaimana yang terdapat pada tabel 3.7.

Tabel 3. 7 *Split* Data

Model	**Data *Train***	**Data *Test***
A	80%	20%
B	70%	30%

Dalam setiap model akan diterapkan *Hyperparameter* dengan detail *Hyperparameter* yang digunakan dapat dilihat pada tabel 3.8 berikut.

Tabel 3. 8 Detail *Hyperparameter*

Hyperparameter	**Nilai**
n_estimators	200
max_depth	20
bootstrap	*False*
random_state	42

Dengan masing-masing model yang terdapat dalam tabel 3.7 akan dilakukan perbandingan uji coba berdasarkan penggunaan normalisasi dan tanpa normalisasi. Dalam masing-masing pengujian tersebut akan dikombinasikan dengan penggunaan *SMOTE* dan seleksi fitur. Sehingga akan dilakukan pengujian sebanyak 8 pengujian pada masing-masing model *split data* 80%:20% dan model *split data* 70%:30%, dengan rincian masing-masing pengujian sebagai berikut:

1. Uji coba tanpa normalisasi, tanpa *SMOTE*, dan tanpa Seleksi Fitur
2. Uji coba tanpa normalisasi, tanpa Seleksi Fitur, namun menggunakan *SMOTE*
3. Uji coba tanpa normalisasi, tanpa *SMOTE*, namun menggunakan seleksi fitur

4. Uji coba tanpa normalisasi, namun menggunakan *SMOTE* dan seleksi fitur
5. Uji coba menggunakan normalisasi, tanpa *SMOTE*, dan tanpa Seleksi Fitur
6. Uji coba menggunakan normalisasi, tanpa Seleksi Fitur, namun menggunakan *SMOTE*
7. Uji coba menggunakan normalisasi, tanpa *SMOTE*, namun menggunakan seleksi fitur
8. Uji coba menggunakan normalisasi, *SMOTE* dan seleksi fitur

BAB IV

UJI COBA DAN PEMBAHASAN

4.1 Pengujian tanpa normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

Pada pengujian ini tidak dilakukan tahap normalisasi, dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split data* yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.1.1 Model A

Pada pengujian ini tidak dilakukan tahap normalisasi sehingga setelah tahap preprocessing selesai akan dilanjutkan dengan pembentukan model dengan algoritma *Random Forest Classifier* dengan perbandingan data *train* dan data *test* sebanyak 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang tertera pada tabel 4.1. Dalam tabel tersebut ditampilkan nilai TP (*True Positive*) sebesar 5889 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 972 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 835 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3320 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 1 *Confusion Matrix* Model A tanpa normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

TP	FP	FN	TN
5889	972	835	3320

Berdasarkan hasil *Confusion Matrix* pada tabel 4.1 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.2.

Tabel 4. 2 *Performance* Model A tanpa normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	84%	88%	86%	87%
1	84%	77%	80%	79%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.1.2 Model B

Pada pengujian ini tidak dilakukan tahap normalisasi sehingga setelah tahap preprocessing selesai akan dilanjutkan dengan pembentukan model dengan algoritma *Random Forest Classifier* dengan perbandingan data *train* dan data *test* sebanyak 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris data dan data *testing* yang digunakan sebanyak 16523 baris data. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang tertera pada tabel 4.3. Dalam tabel 4.3 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8873 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1684 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 1280 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 4871 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 3 *Confusion Matrix* Model B tanpa normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

TP	**FP**	**FN**	**TN**
8873	1684	1280	4871

Berdasarkan hasil *Confusion Matrix* pada tabel 4.3 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.4.

Tabel 4. 4 *Performance* Model B tanpa normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

Class	Accuracy	Precision	Recall	F1-Score
0	82%	87%	84%	86%
1	82%	74%	79%	77%

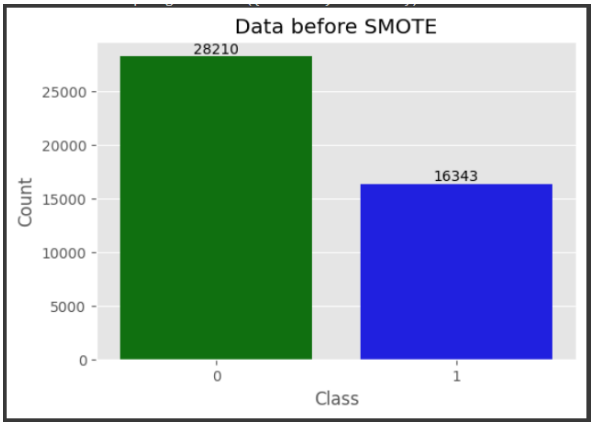
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.2 Pengujian dengan *SMOTE*, tanpa normalisasi dan tanpa Seleksi Fitur

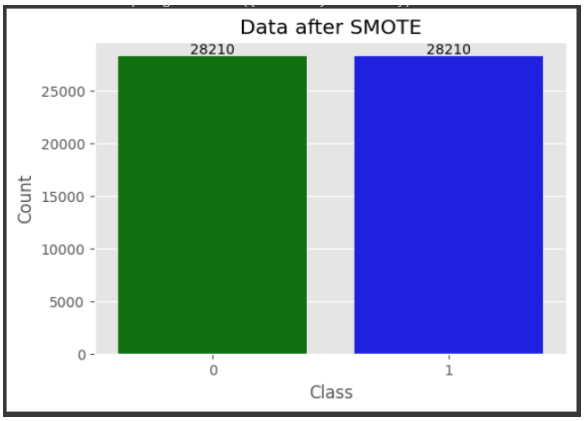
Pada pengujian ini tidak dilakukan tahap normalisasi, namun dilakukan implementasi *SMOTE* untuk mengatasi jumlah data target yang tidak seimbang. Dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split data* yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.2.1 Model A

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE,* jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.1 dan gambar 4.2.



Gambar 4. 1 Data Model A sebelum *SMOTE*



Gambar 4. 2 Data Model A setelah *SMOTE*

Setelah dilakukan proses penyeimbangan data dengan *SMOTE*, dilanjutkan dengan pembentukan model sehingga menghasilkan *Confusion Matrix* yang terdapat pada tabel 4.5. Dalam tabel 4.5 tersebut ditampilkan nilai TP (*True Positive*) sebesar 5602 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1425 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 615 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3497 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 5 *Confusion Matrix* Model A *SMOTE*, tanpa normalisasi dan tanpa seleksi fitur

TP	FP	FN	TN
5602	1425	615	3497

Berdasarkan hasil *Confusion Matrix* pada tabel 4.5 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.6.

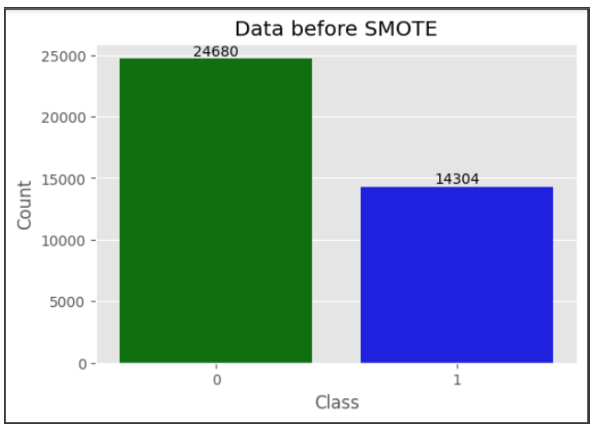
Tabel 4. 6 *Performance* Model A *SMOTE*, tanpa normalisasi dan tanpa seleksi fitur

Class	Accuracy	Precision	Recall	F1-Score
0	82%	90%	80%	85%
1	82%	71%	85%	77%

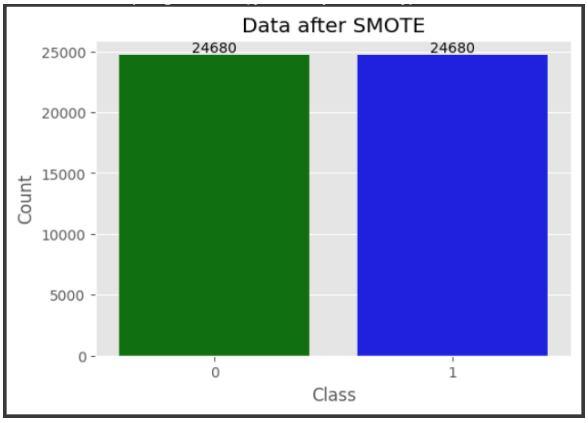
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.2.2 Model B

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris data dan data *testing* yang digunakan sebanyak 16523 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE*, jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.3 dan gambar 4.4.



Gambar 4. 3 Data Model B sebelum *SMOTE*



Gambar 4. 4 Data Model B setelah *SMOTE*

Setelah dilakukan proses penyeimbangan data dengan *SMOTE*, dilanjutkan dengan pembentukan model sehingga menghasilkan *Confusion Matrix* yang terdapat pada tabel 4.7. Dalam tabel 4.7 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8400 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 2157 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 968 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 5183 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 7 *Confusion Matrix* Model B *SMOTE*, tanpa normalisasi dan tanpa seleksi fitur

TP	**FP**	**FN**	**TN**
8400	2157	968	5183

Berdasarkan hasil *Confusion Matrix* pada tabel 4.7 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.8.

Tabel 4. 8 *Performance* Model B *SMOTE*, tanpa normalisasi dan tanpa seleksi fitur

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	81%	90%	80%	84%
1	81%	71%	84%	77%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.3 Pengujian tanpa normalisasi, tanpa *SMOTE* dan implementasi Seleksi Fitur

Pada pengujian ini tidak dilakukan tahap normalisasi, namun dilakukan implementasi seleksi fitur untuk memilih 13 fitur teratas yang terdapat dalam dataset yang digunakan. Dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split data* yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.3.1 Model A

Pada pengujian ini dilakukan split data dengan perbandingan data train dan data test sebesar 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Dengan dilanjutkan dengan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.9.

Tabel 4. 9 *Rank* 13 fitur teratas Model A tanpa normalisasi

Rank	**Fitur**
1	*Gender*
2	*Hemoglobin*
3	*Gtp*
4	*Height(cm)*
5	*Triglyceride*
6	*Waist(cm)*
7	*LDL*
8	*ALT*
9	*Cholesterol*
10	*HDL*
11	*Fasting blood sugar*
12	*Systolic*

Rank	**Fitur**
13	*Age*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.10. Dalam tabel 4.10 tersebut ditampilkan nilai TP (*True Positive*) sebesar 5938 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1089 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 843 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3269 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 10 *Confusion Matrix* Model A seleksi fitur, tanpa normalisasi dan tanpa *SMOTE*

TP	**FP**	**FN**	**TN**
5938	1089	843	3269

Berdasarkan hasil *Confusion Matrix* pada tabel 4.10 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.11.

Tabel 4. 11 *Performance* Model A seleksi fitur, tanpa normalisasi dan tanpa *SMOTE*

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	83%	88%	85%	86%
1	83%	75%	79%	77%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.3.2 Model B

Pada pengujian ini dilakukan split data dengan perbandingan data train dan data test sebesar 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris

data dan data *testing* yang digunakan sebanyak 16523 baris data. Dengan dilanjutkan dengan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.12

Tabel 4. 12 *Rank* 13 fitur teratas Model B tanpa normalisasi

Rank	**Fitur**
1	*Gender*
2	*Gtp*
3	*Hemoglobin*
4	*Height(cm)*
5	*Triglyceride*
6	*Waist(cm)*
7	*LDL*
8	*ALT*
9	*Cholesterol*
10	*Fasting blood sugar*
11	*HDL*
12	*Systolic*
13	*AST*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.13. Dalam tabel 4.13 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8850 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1707 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 1422 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 4729 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 13 *Confusion Matrix* Model B seleksi fitur, tanpa normalisasi dan tanpa *SMOTE*

TP	**FP**	**FN**	**TN**
8850	1707	1422	4729

Berdasarkan hasil *Confusion Matrix* pada tabel 4.13 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.14.

Tabel 4. 14 *Performance* Model B seleksi fitur, tanpa normalisasi dan tanpa *SMOTE*

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	81%	86%	84%	85%
1	81%	73%	77%	75%

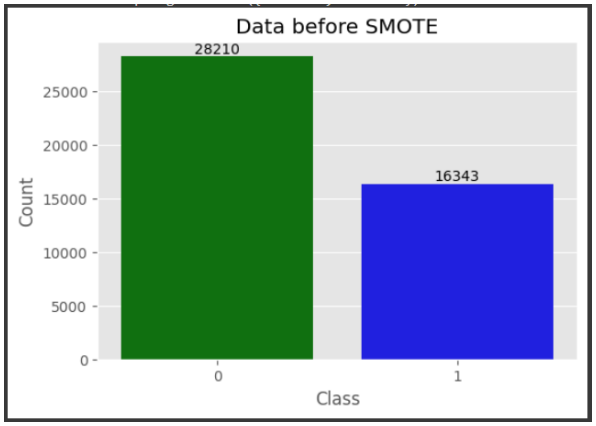
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.4 Pengujian tanpa normalisasi, implementasi *SMOTE* dan seleksi fitur

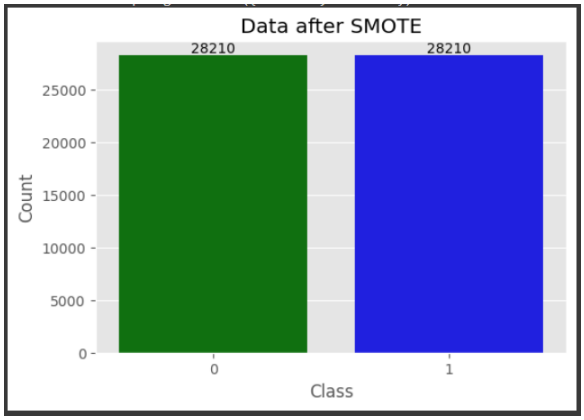
Pada pengujian ini tidak dilakukan tahap normalisasi, namun dilakukan implementasi *SMOTE* untuk mengatasi jumlah data target yang tidak seimbang dan implementasi seleksi fitur untuk memilih 13 fitur teratas yang terdapat dalam dataset yang digunakan. Dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan split data yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.4.1 Model A

Pada pengujian ini dilakukan split data dengan perbandingan data train dan data test sebesar 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE*, jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.5 dan gambar 4.6.



Gambar 4. 5 Data Model A sebelum *SMOTE*



Gambar 4. 6 Data Model A setelah *SMOTE*

Kemudian, tahap berikutnya dijalankan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.15.

Tabel 4. 15 *Rank* 13 fitur teratas Model A *SMOTE* tanpa normalisasi

Rank	**Fitur**
1	*Gender*
2	*Height(cm)*
3	*Hemoglobin*
4	*Gtp*
5	*Serum creatinine*
6	*Age*
7	*Triglyceride*
8	*Weight(kg)*
9	*Waist(cm)*
10	*ALT*
11	*LDL*
12	*HDL*

Rank	**Fitur**
13	*Cholesterol*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.16. Dalam tabel 4.16 tersebut ditampilkan nilai TP (*True Positive*) sebesar 5650 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1377 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 668 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3444 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 16 *Confusion Matrix* Model A *SMOTE*, seleksi fitur dan tanpa normalisasi

TP	**FP**	**FN**	**TN**
5650	1377	668	3444

Berdasarkan hasil *Confusion Matrix* pada tabel 4.16 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.17.

Tabel 4. 17 *Performance* Model A *SMOTE*, seleksi fitur dan tanpa normalisasi

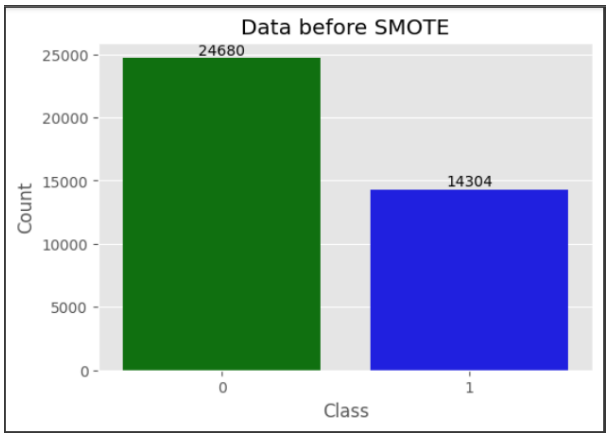
Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	82%	86%	84%	85%
1	82%	73%	77%	75%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

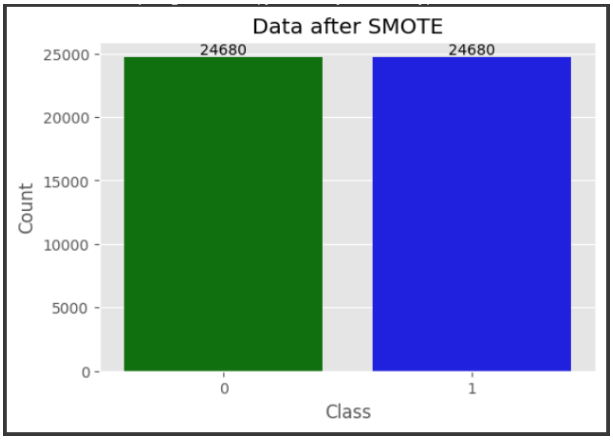
4.4.2 Model B

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris

data dan data *testing* yang digunakan sebanyak 16523 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE*, jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.7 dan gambar 4.8.



Gambar 4. 7 Data Model B sebelum *SMOTE*



Gambar 4. 8 Data Model B setelah *SMOTE*

Kemudian, tahap berikutnya dijalankan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.18.

Tabel 4. 18 *Rank* 13 fitur teratas Model B *SMOTE tanpa normalisasi*

Rank	**Fitur**
1	*Gender*
2	*Height(cm)*

Rank	**Fitur**
3	*Hemoglobin*
4	*Gtp*
5	*Triglyceride*
6	*Serum creatinine*
7	*Age*
8	*Weight(kg)*
9	*Waist(cm)*
10	*ALT*
11	*HDL*
12	*LDL*
13	*Cholesterol*

Setelah pemiilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.19. Dalam tabel 4.19 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8449 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 2108 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 1102 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 5049 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 19 *Confusion Matrix* Model B *SMOTE*, seleksi fitur dan tanpa normalisasi

TP	**FP**	**FN**	**TN**
8449	2108	1102	5049

Berdasarkan hasil *Confusion Matrix* pada tabel 4.19 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.20.

Tabel 4. 20 *Performance* Model B *SMOTE*, seleksi fitur dan tanpa normalisasi

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	81%	86%	84%	85%
1	81%	73%	77%	75%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.5 Pengujian dengan normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

Pada pengujian ini dilakukan tahap normalisasi, dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split* data yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.5.1 Model A

Pada pengujian ini dilakukan tahap normalisasi dalam tahap *preprocessing*-nya. Kemudian dilanjutkan dengan pembentukan model dengan algoritma *Random Forest Classifier* dengan perbandingan data *train* dan data *test* sebanyak 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang tertera pada tabel 4.21. Dalam tabel 4.21 tersebut ditampilkan nilai TP (*True Positive*) sebesar 5889 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 972 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 835 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3320 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 21 *Confusion Matrix* Model A dengan normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

TP	**FP**	**FN**	**TN**
5889	972	835	3320

Berdasarkan hasil *Confusion Matrix* pada tabel 4.21 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.22.

Tabel 4. 22 *Performance* Model A dengan normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	84%	88%	86%	87%
1	84%	77%	80%	79%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.5.2 Model B

Pada pengujian ini dilakukan tahap normalisasi dalam tahap preprocessing-nya. Kemudian dilanjutkan dengan pembentukan model dengan algoritma *Random Forest Classifier* dengan perbandingan data *train* dan data *test* sebanyak 80%:20%. Data *training* yang digunakan sebanyak 38.553 baris data dan data *testing* yang digunakan sebanyak 16523 baris data. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang tertera pada tabel 4.23. Dalam tabel 4.23 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8853 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1538 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 1315 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 4817 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 23 *Confusion Matrix* Model B dengan normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

TP	**FP**	**FN**	**TN**
8853	1538	1315	4817

Berdasarkan hasil *Confusion Matrix* pada tabel 4.23 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.24.

Tabel 4. 24 *Performance* Model B dengan normalisasi, tanpa *SMOTE* dan tanpa Seleksi Fitur

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	83%	87%	85%	86%
1	83%	76%	79%	77%

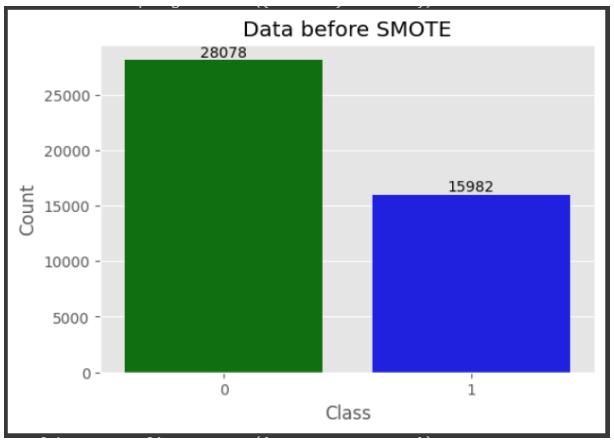
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.6 Pengujian dengan normalisasi dan implementasi *SMOTE* tanpa Seleksi Fitur

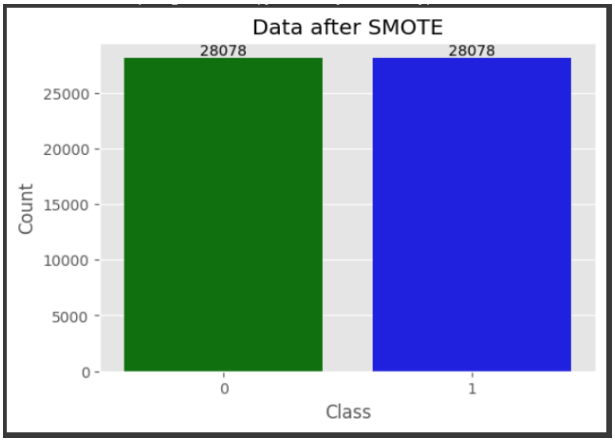
Pada pengujian ini dilakukan tahap normalisasi, dan dilanjutkan dengan implementasi *SMOTE* untuk mengatasi jumlah data target yang tidak seimbang. Dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split* data yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.6.1 Model A

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE,* jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.9 dan gambar 4.10.



Gambar 4. 9 Data Model A dengan normalisasi sebelum *SMOTE*



Gambar 4. 10 Data Model A dengan normalisasi setelah *SMOTE*

Setelah dilakukan proses penyeimbangan data dengan *SMOTE*, dilanjutkan dengan pembentukan model sehingga menghasilkan *Confusion Matrix* yang terdapat pada tabel 4.25. Dalam tabel 4.25 tersebut ditampilkan nilai TP (*True Positive*) sebesar 5581 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1280 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 627 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3528 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 25 *Confusion Matrix* Model A normalisasi *SMOTE* dan tanpa seleksi fitur

TP	FP	FN	TN
5581	1280	627	3528

Berdasarkan hasil *Confusion Matrix* pada tabel 4.25 diperoleh perhitungan *Accuracy, precision, recall* dan *f1-score* yang terdapat pada tabel 4.26.

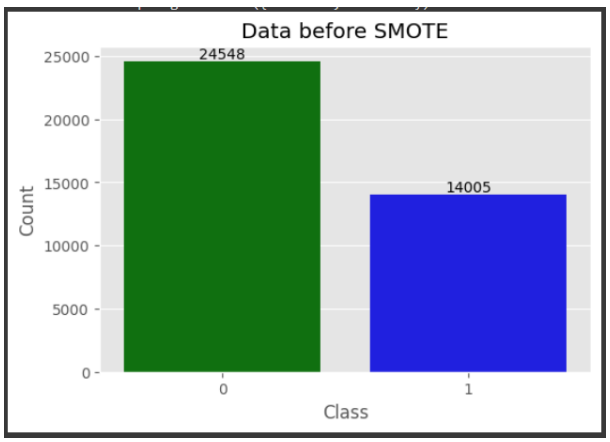
Tabel 4. 26 *Performance* Model A normalisasi *SMOTE* dan tanpa seleksi fitur

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	83%	90%	81%	85%
1	83%	73%	85%	79%

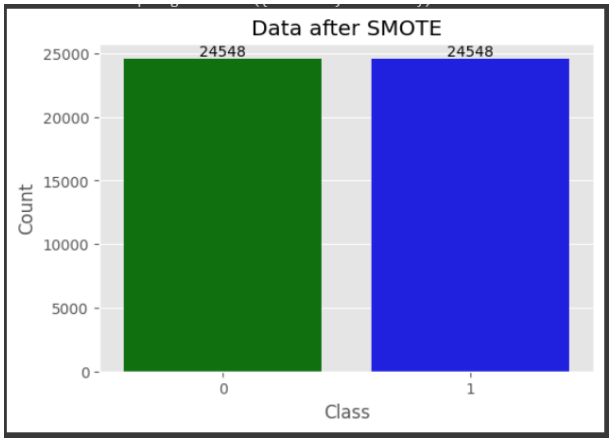
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.6.2 Model B

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris data dan data *testing* yang digunakan sebanyak 16523 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE*, jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.11 dan gambar 4.12.



Gambar 4. 11 Data Model B dengan normalisasi sebelum *SMOTE*



Gambar 4. 12 Data Model B dengan normalisasi setelah *SMOTE*

Setelah dilakukan proses penyeimbangan data dengan *SMOTE*, dilanjutkan dengan pembentukan model sehingga hasil *Confusion Matrix* yang terdapat pada tabel 4.27. Dalam tabel 4.27 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8873 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1684 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 1280 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 4871 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 27 *Confusion Matrix* Model B normalisasi *SMOTE* dan tanpa seleksi fitur

TP	**FP**	**FN**	**TN**
8377	2014	969	5163

Berdasarkan hasil *Confusion Matrix* pada tabel 4.27 diperoleh perhitungan *Accuracy, precision, recall* dan *f1-score* yang terdapat pada tabel 4.28.

Tabel 4. 28 *Performance* Model B normalisasi *SMOTE* dan tanpa seleksi fitur

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	82%	90%	81%	85%
1	82%	72%	84%	78%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.7 Pengujian dengan normalisasi dan implementasi Seleksi fitur

Pada pengujian ini dilakukan tahap normalisasi, dan dilanjutkan dengan implementasi seleksi fitur untuk memilih 13 fitur teratas yang terdapat dalam *dataset* yang digunakan. Dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split* data yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.7.1 Model A

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Dengan dilanjutkan dengan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.29.

Tabel 4. 29 *Rank* 13 fitur teratas Model A dengan normalisasi

Rank	**Fitur**
1	*Gender*
2	*Gtp*
3	*Hemoglobin*
4	*Height(cm)*
5	*Triglyceride*
6	*Waist(cm)*
7	*LDL*
8	*ALT*
9	*Cholesterol*
10	*HDL*
11	*Fasting blood sugar*
12	*Age*
13	*Systolic*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.30. Dalam tabel 4.30 ditampilkan nilai TP (*True Positive*) sebesar 5894 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 967 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 867 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3288 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 30 *Confusion Matrix* Model A seleksi fitur, normalisasi dan tanpa *SMOTE*

TP	**FP**	**FN**	**TN**
5894	967	867	3288

Berdasarkan hasil *Confusion Matrix* pada tabel 4.30 diperoleh perhitungan *Accuracy, precision, recall* dan *f1-score* yang terdapat pada tabel 4.31.

Tabel 4. 31 *Performance* Model A seleksi fitur, normalisasi dan tanpa *SMOTE*

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	83%	87%	86%	87%
1	83%	77%	79%	78%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.7.2 Model B

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris data dan data *testing* yang digunakan sebanyak 16523 baris data. Dengan

dilanjutkan dengan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.32.

Tabel 4. 32 *Rank* 13 fitur teratas Model B dengan normalisasi

Rank	**Fitur**
1	*Gender*
2	*Gtp*
3	*Hemoglobin*
4	*Height(cm)*
5	*Triglyceride*
6	*Waist(cm)*
7	*LDL*
8	*ALT*
9	*Cholesterol*
10	*HDL*
11	*Fasting blood sugar*
12	*Systolic*
13	*Age*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.33. Dalam tabel 4.33 tersebut ditampilkan nilai TP (*True Positive*) sebesar 8853 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1538 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 1315 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 4817 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 33 *Confusion Matrix* Model B seleksi fitur, normalisasi dan tanpa *SMOTE*

TP	**FP**	**FN**	**TN**
8831	1560	1382	4750

Berdasarkan hasil *Confusion Matrix* pada tabel 4.33 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.34.

Tabel 4. 34 *Performance* Model B seleksi fitur, normalisasi dan tanpa *SMOTE*

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	82%	86%	85%	86%
1	82%	75%	77%	76%

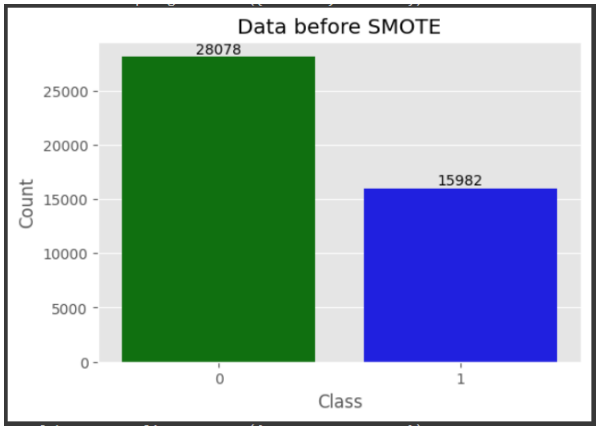
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.8 Pengujian dengan normalisasi, implementasi *SMOTE* dan seleksi fitur

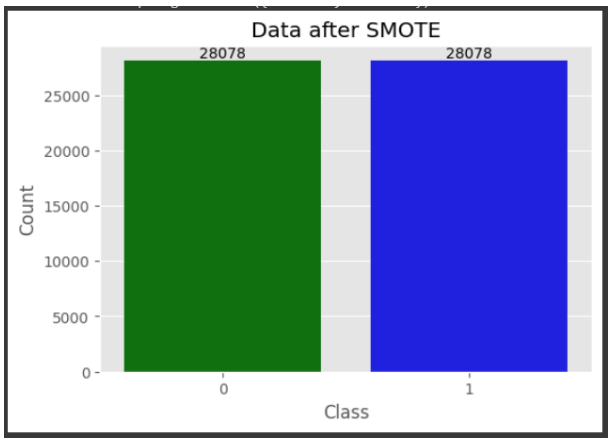
Pada pengujian ini dilakukan tahap normalisasi, kemudian dilanjutkan dengan implementasi *SMOTE* untuk mengatasi jumlah data target yang tidak seimbang dan tahap berikutnya yakni proses seleksi fitur untuk memilih 13 fitur teratas yang terdapat dalam dataset yang digunakan. Dalam jenis pengujian ini akan dibuat dua model berbeda berdasarkan *split* data yang digunakan, yaitu Model A dengan data *train* dan data *test* sebanyak 80%:20% sedangkan Model B data *train* dan data *test* sebanyak 70%:30%.

4.8.1 Model A

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 80%:20%. Data *training* yang digunakan sebanyak 44.060 baris data dan data *testing* yang digunakan sebanyak 11016 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE,* jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.13 dan gambar 4.14.



Gambar 4. 13 Data Model A dengan normalisasi sebelum *SMOTE*



Gambar 4. 14 Data Model A dengan normalisasi setelah *SMOTE*

Kemudian, tahap berikutnya dijalankan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.35.

Tabel 4. 35 *Rank* 13 fitur teratas Model A *SMOTE* dengan normalisasi

Rank	**Fitur**
1	*Gender*
2	*Height(cm)*
3	*Hemoglobin*
4	*Gtp*
5	*Triglyceride*
6	*Serum creatinine*
7	*Age*
8	*Weight(kg)*
9	*Waist(cm)*
10	*ALT*
11	*LDL*

Rank	**Fitur**
12	*HDL*
13	*Cholesterol*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.36. Dalam tabel 4.36 tersebut ditampilkan nilai TP (*True Positive*) sebesar 5602 yang memprediksi bukan perokok dan benar bukan perokok, FP (*False Positive*) bernilai 1259 yang memprediksi perokok namun sebenarnya bukan perokok, FN (*False Negative*) bernilai 661 yang memprediksi perokok dan namun sebenarnya bukan perokok, serta nilai TN (*True Negative*) sebesar 3494 yang memprediksi perokok dan sebenarnya perokok.

Tabel 4. 36 *Confusion Matrix* Model A *SMOTE*, seleksi fitur dan normalisasi

TP	**FP**	**FN**	**TN**
5602	1259	661	3494

Berdasarkan hasil *Confusion Matrix* pada tabel 4.36 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.37.

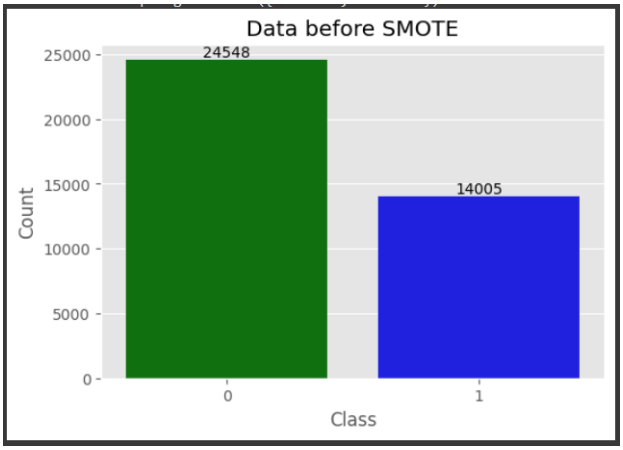
Tabel 4. 37 *Performance* Model A *SMOTE*, seleksi fitur dan normalisasi

Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	83%	89%	82%	85%
1	83%	74%	84%	78%

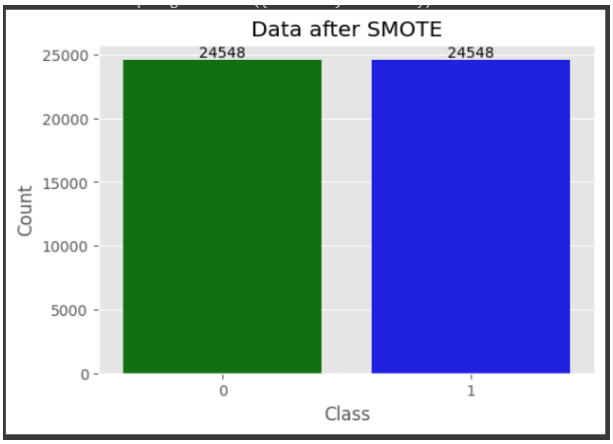
Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.8.2 Model B

Pada pengujian ini dilakukan *split* data dengan perbandingan data *train* dan data *test* sebesar 70%:30%. Data *training* yang digunakan sebanyak 38.553 baris data dan data *testing* yang digunakan sebanyak 16523 baris data. Dengan dilanjutkan dengan proses penyeimbangan data menggunakan *SMOTE*, jumlah data target sebelum dan sesudah dilakukan proses *SMOTE* ditunjukkan pada gambar 4.15 dan gambar 4.16.



Gambar 4. 15 Data Model B dengan normalisasi sebelum *SMOTE*



Gambar 4. 16 Data Model B dengan normalisasi setelah *SMOTE*

Kemudian, tahap berikutnya dijalankan proses seleksi fitur menggunakan teknik *feature importances*. Sebanyak 13 fitur teratas yang terpilih terdapat pada tabel 4.38.

Tabel 4. 38 *Rank* 13 fitur teratas Model B *SMOTE* dengan normalisasi

Rank	**Fitur**
1	Gender
2	Height(cm)
3	Hemoglobin
4	Gtp
5	Triglyceride
6	Serum creatinine
7	Age
8	Weight(kg)
9	Waist(cm)
10	ALT
11	HDL
12	LDL
13	*Cholesterol*

Setelah pemilihan fitur tersebut, tahap berikutnya adalah pembentukan model dengan fitur yang telah diseleksi sebelumnya. Pada proses tersebut, diperoleh hasil *Confusion Matrix* yang terdapat pada tabel 4.39.

Tabel 4. 39 *Confusion Matrix* Model B *SMOTE*, seleksi fitur dan normalisasi

TP	**FP**	**FN**	**TN**
8379	2012	1065	5067

Berdasarkan hasil *Confusion Matrix* pada tabel 4.39 diperoleh perhitungan *Accuracy, precision, recall dan f1-score* yang terdapat pada tabel 4.40.

Tabel 4. 40 *Performance* Model B *SMOTE*, seleksi fitur dan normalisasi

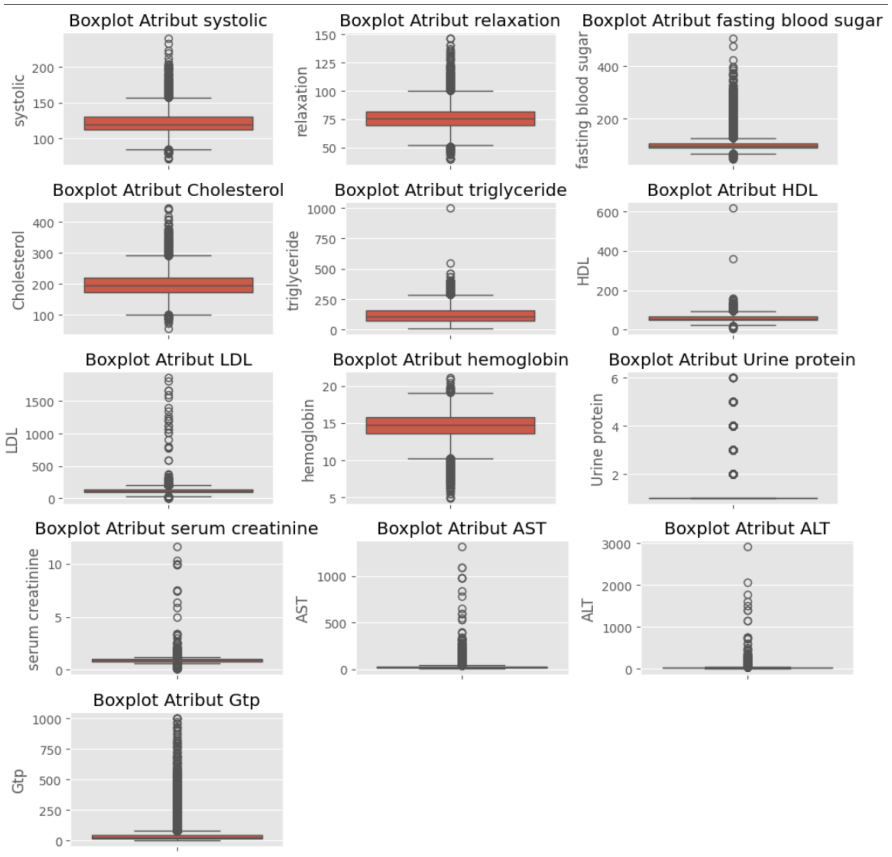
Class	***Accuracy***	***Precision***	***Recall***	***F1-Score***
0	81%	89%	81%	84%
1	81%	72%	83%	77%

Class 0 merupakan inisialisasi untuk kategori bukan perokok, sedangkan *Class* 1 adalah inisialisasi untuk kategori perokok.

4.9 Pembahasan

Berdasarkan hasil dari pengujian yang telah dilakukan pada sub-bab sebelumnya, pada sub bab ini akan dijelaskan lebih detail tentang hasil uji coba yang telah dilakukan. *Dataset* yang digunakan untuk penelitian ini adalah *dataset* yang diperoleh dari salah satu Perusahaan Asuransi Kesehatan Nasional milik Republik Korea Selatan. Dalam dataset tersebut berisi 55692 baris data yang berasal dari hasil pemeriksaan pasien yang tinggal di negara tersebut.

Dalam tahap preprocessing dilakukan pencarian *outlier* dengan visualisasi data menggunakan *EDA* yang divisualisasikan pada gambar 4.17. Kemudian, *outlier* tersebut akan dihapus untuk pembersihan data. Dalam dataset tersebut terdapat sebanyak 616 baris data yang memiliki *outlier*.



Gambar 4. 17 *Boxplot* fitur-fitur yang memiliki *outlier*

Setelah dilakukan penghapusan *outlier*, tahap berikutnya yaitu *encoding categorical data* untuk mengubah data kategorikal menjadi data angka. Selanjutnya dilakukan proses normalisasi sesuai dengan pengujian yang terdapat pada skenario pengujian. Berdasarkan pengujian yang telah dilakukan, diperoleh hasil yang diringkas dalam tabel 4.41 untuk pengujian tanpa normalisasi dan tabel 4.42 untuk pengujian dengan normalisasi. Dengan perbandingan data pada model A sebanyak 80% data *train* 20% data *test*, serta perbandingan data pada model B sebanyak 70% data *train* 30% data *test*.

Tabel 4. 41 Hasil pengujian Model tanpa normalisasi

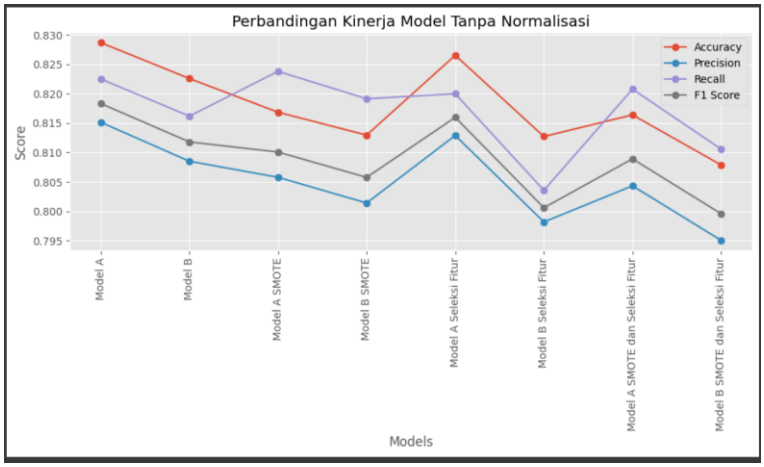
Tanpa Normalisasi					
Model	**Implementasi**	***Accuracy***	***Precision***	***Recall***	***F1-Score***
A	Tanpa *SMOTE* dan Tanpa Seleksi Fitur	82,87%	84,62%	87,79%	86,17%
B		82,26%	84,05%	87,39%	85,69%
A	*SMOTE* dan Tanpa Seleksi Fitur	81,69%	79,72%	90,11%	84,60%
B		81,30%	79,57%	89,67%	84,32%
A	Tanpa *SMOTE* namun menggunakan Seleksi fitur	82,66%	84,50%	87,57%	86,01%
B		81,27%	83,83%	86,16%	84,98%
A	*SMOTE* dan Seleksi fitur	81,64%	80,40%	89,43%	84,68%
B		80,79%	80,03%	88,46%	84,04%

Tabel 4. 42 Hasil pengujian Model dengan normalisasi

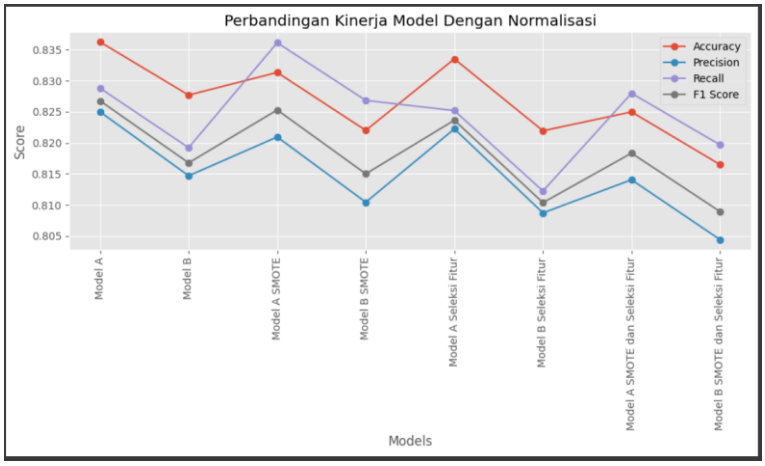
Normalisasi					
Model	**Implementasi**	***Accuracy***	***Precision***	***Recall***	***F1-Score***
A	Tanpa *SMOTE* dan Tanpa Seleksi Fitur	83,62%	85,91%	87,56%	86,73%
B		82,77%	85,21%	87,11%	86,15%
A	*SMOTE* dan Tanpa Seleksi Fitur	83,13%	81,66%	90,33%	85,78%
B		82,20%	80,81%	89,87%	85,10%
A	Tanpa *SMOTE* namun menggunakan Seleksi fitur	83,35%	85,91%	87,18%	86,54%
B		82,19%	84,99%	86,47%	85,72%
A	*SMOTE* dan Seleksi fitur	82,50%	81,58%	89,39%	85,31%
B		81,65%	80,73%	89,06%	84,69%

Berdasarkan hasil pengujian yang terdapat pada tabel 4.41 dibuat grafik yang direpresentasikan pada gambar 4.18, pada grafik tersebut ditunjukkan

perbandingan kinerja model yang dibuat tanpa normalisasi. Sedangkan hasil pengujian tabel 4.42 dibuat grafik yang direpresentasikan pada gambar 4.19 yang menunjukkan perbandingan kinerja model menggunakan normalisasi. Hasil grafik yang fluktuatif dari kedua grafik menginformasikan bahwa ketika dilakukan *SMOTE* hasil *Accuracy, precision* dan *f1-score* menurun namun *recall* meningkat. Dengan meningkatnya *Recall* menunjukkan bahwa model menjadi lebih baik dalam mendeteksi kasus-kasus positif minoritas. Salah satu hasil Tree dari algoritma *Random Forest* tersebut dapat dilihat pada **Lampiran 1**.



Gambar 4. 18 Grafik *Performance* Model tanpa Normalisasi



Gambar 4. 19 Grafik *Performance* Model dengan Normalisasi

Pada grafik yang ditunjukkan gambar 4.18 dan 4.19, Model yang melakukan pengujian menggunakan *SMOTE* dan Seleksi fitur memiliki nilai *Accuracy, Precision, Recall*, dan *F1-Score* lebih rendah daripada Model yang melakukan pengujian dengan tanpa *SMOTE* dan tanpa Seleksi Fitur. Hal tersebut dapat terjadi karena telah banyak proses yang dilalui yaitu *SMOTE* dan Seleksi Fitur. Salah satu faktor yang menyebabkan hal tersebut terjadi karena penggunaan *SMOTE* yang membuat data sintetis dengan jumlah data yang ditambahkan terlalu besar dibandingkan data asli, serta efek dari Seleksi Fitur yaitu apabila fitur-fitur yang dipilih tidak mencakup informasi yang relevan, sehingga dapat mengurangi fitur-fitur yang penting.

Model yang tidak mengimplementasikan *SMOTE* memiliki nilai *Accuracy, Precision, Recall*, dan *F1-Score* lebih tinggi daripada Model yang mengimplementasikan *SMOTE*. Hal tersebut terjadi karena Model yang mengimplementasikan *SMOTE* memiliki data sintetis yang ditambahkan untuk menyeimbangkan data. Model yang mengimplementasikan *SMOTE* juga memiliki nilai *Precision* yang lebih tinggi dari pada *Accuracy, Recall* dan *F1-score* sehingga Model yang mengimplementasikan *SMOTE* lebih mampu dalam melakukan identifikasi sebagian besar kasus positif sebenarnya untuk mencegah terjadinya identifikasi yang melewatkan kasus poditif sebenarnya.

Model yang mengimplementasikan Seleksi Fitur memiliki nilai *Accuracy, Peecision, Recall* dan *F1-Score* yang lebih rendah daripada Model yang tidak mengimplementasikan Seleksi Fitur. Hal ini dapat terjadi karena Model yang menggunakan Seleksi Fitur akan membuat model dengan menggunakan sebagian

Fitur *dataset* yang ada, sehingga memungkinkan terjadinya beberapa fitur penting lainnya terbuang atau hilang.

Tabel 4. 43 Hasil Pengujian terbaik

Normalisasi					
Model	**Implementasi**	***Accuracy***	***Precision***	***Recall***	***F1-Score***
A (80%:20%)	*SMOTE* dan Seleksi fitur	82,50%	81,58%	89,39%	85,31%
B (70%:30%)		81,65%	80,73%	89,06%	84,69%

Dengan beberapa hasil pengujian tersebut, model terbaik diperoleh dari hasil pengujian dengan normalisasi dan *SMOTE* serta Seleksi Fitur. Dengan Model A dengan perbandingan data *train* 80% dan data *test* 20% yang mengimplementasikan normalisasi dan *SMOTE* serta seleksi fitur menghasilkan nilai *Accuracy* 82,50%, *precision* 81,58%, *recall* 89,39% dan *f1-score* 85,31% dan Model B dengan perbandingan data *train* 70% dan data *test* 30% menghasilkan nilai *Accuracy* 81,65%, *precision* 80,73%, *recall* 89,06%, dan *f1-score* 84,69%. Hal tersebut terjadi karena model yang memiliki data yang seimbang dapat terhindar dari hasil *Accuracy* metrik yang menyesatkan, serta proses seleksi fitur yang digunakan bisa menjadikan model tersebut lebih *interpretable* dan lebih cepat dalam melakukan prediksi dibandingkan dengan model yang menggunakan semua fitur.

Berdasarkan nilai akurasi yang telah diperoleh menggunakan pengujian terbaik, disebutkan oleh (Subarkah et al., 2022) bahwa nilai *Accuracy* tergolong baik saat mencapai nilai 80%-90%, kategori klasifikasi sangat baik apabila mencapai nilai 90%-100%, cukup baik pada nilai 70%-80%, 60%-70% termasuk klasifikasi kurang baik dan klasifikasi gagal saat nilai dalam rentang 50%-60%. Sehingga dapat diketahui bahwa model klasifikasi perokok berdasarkan kondisi

tubuh dengan metode *Random Forest* yang dibuat dalam penelitian ini menghasilkan klasifikasi yang baik dan dapat menambah wawasan tentang keterkaitan perilaku merokok terhadap kondisi tubuh.

4.10 Integrasi Islam

4.10.1 Muamalah mu'Allah

Dalam upaya meningkatkan wawasan tentang akibat merokok terhadap kesehatan tubuh. Hal tersebut merupakan upaya dalam melakukan penjagaan diri dari penyakit dengan memperhatikan makanan yang dikonsumsi, sebagaimana yang telah terdapat dalam Q.S. al-Baqarah ayat 168 yang berbunyi:

يَا أَيُّهَا النَّاسُ كُلُوا مِمَّا فِي الْأَرْضِ حَلَالًا طَيِّبًا وَلَا تَتَّبِعُوا خُطُوَاتِ الشَّيْطَانِ ، إِنَّهُ لَكُمْ عَدُوٌّ مُبِينٌ

"Wahai manusia, makanlah yang halal lagi baik dari apa yang terdapat di bumi dan janganlah kamu mengikuti langkah-langkah syaithan; karena sesungguhnya syaithan adalah musuh yang nyata bagimu" (Q.S. Al-Baqarah: 168)

Dalam Tafsir Ibnu Katsir dijelaskan bahwa dalam ayat tersebut Allah memperbolehkan manusia untuk memakan segala yang ada di muka bumi, yaitu makanan yang halal, baik dan bermanfaat bagi dirinya serta tidak membahayakan bagi tubuh dan akal pikirannya. Salah satu perilaku yang membahayakan bagi tubuh dan akal pikiran adalah aktivitas merokok. Perilaku merokok yang dilakukan oleh manusia termasuk dalam perbuatan yang bertentangan dengan perintah Allah untuk makan makanan yang halal dan baik. Dalam hal tersebut, apabila terdapat sesuatu yang tidak baik masuk ke dalam tubuh manusia akan dapat memberikan efek yang buruk. Salah satunya yaitu dapat mempengaruhi kinerja organ dalam tubuh yang dapat menjadi asal mula penyakit.

Berdasarkan efek buruk tersebut, mayoritas ulama dalam mahdzab empat memberikan pendapat dengan melarang rokok dengan keras. Menurut mahdzab empat, diharamkan merokok karena rokok tersebut dapat membuka jalan agar tubuh terjangkit berbagai penyakit berbahaya. Namun, ada pendapat pula yang mengatakan bahwa hukum merokok adalah makruh. Hal tersebut didasarkan pada dampak negatif merokok terhadap kesehatan individu dan masyarakat umum. Menurut Sheikh Yusuf al Qaradhawi (Qaradhawi, 1993) asal mula setiap sesuatu adalah mubah, salah satunya yaitu merokok diperbolehkan namun tidak disukai.

4.10.2 Muamalah Mu'Annas

Sebagaimana yang terdapat dalam al Qur'an yang memiliki banyak ayat-ayat yang mengajarkan manusia tentang berbuat kebaikan, salah satunya diajarkan tentang saling tolong menolong dan berbuat kebajikan untuk kepentingan bersama. Salah satu penerapannya dengan memberikan manfaat kepada orang lain berupa informasi tentang kondisi tubuh yang berpengaruh bagi seseorang yang memiliki riwayat merokok, dengan niat yang didasarkan kepada kebaikan dan ketakwaan kepada Allah SWT. sebagaimana yang telah Allah perintahkan kepada hambanya untuk saling tolong menolong dalam surah Al Maidah ayat 2:

وَتَعَاوَنُوْا عَلَى الْبِرِّ وَالتَّقْوٰى وَلَا تَعَاوَنُوا عَلَى الْاِثْمِ وَالْعُدْوَانِ وَاتَّقُوْا اللّٰهَ اِنَّ اللّٰهَ شَدِيْدُ الْعِقَابِ

"Dan tolong-menolonglah kamu dalam mengerjakan kebajikan dan takwa, dan jangan tolong menolong dalam mengerjakan perbuatan dosa dan permusuhan. Bertakwalah kepada Allah, sesungguhnya Allah sangat berat siksaan-Nya", (Q.S. Al-Maidah:2)

Dalam tafsir ibnu katsir dijelaskan dalam ayat tersebut diperintahkan untuk saling membantu dalam hal kebaikan dan dilarang untuk saling membantu dalam perbuatan tercela. Hal tersebut berarti juga dengan larangan untuk menyakiti dan membahayakan keselamatan orang lain. Sebagaimana yang dijelaskan dalam Hadits Arbain tentang larangan membahayakan diri dan orang lain, Dari Abu Sa'ad bin Malik bin Sinan Al Khudry r.a., bahwa Rasulullah SAW bersabda:

لَا ضَرَرَ وَلَا ضِرَارَ

"*Tidak boleh melakukan sesuatu yang berbahaya dan menimbulkan bahaya bagi orang lain*". (Hadits hasan diriwatkan oleh Ibnu Majah, Ad Daruquthni dan lainnya dengan sanad bersambung)

Berdasarkan larangan dalam hadits tersebut yang berbanding lurus dengan pendapat mayoritas ulama dalam mahdzab empat yang mengharamkan merokok. Namun, ada pendapat yang menghukumi merokok sebagai hal yang makruh yang didasarkan pada banyaknya hal buruk yang terkandung didalamnya, tetapi tidak dijelaskan secara terang-terangan dalam Al-Qur'an. Sebagaimana yang terdapat dalam (Qaradhawi, 1993) yang menyatakan bahwa asal mula segala sesuatu hukumnya adalah mubah, namun dalam masalah ini merokok diperbolehkan namun tidak disukai.

4.10.3 Muamalah Mu'Alam

وَإِذَا تَوَلَّى سَعَى فِي الأَرْضِ لِيُفْسِدَ فِيْهَا وَيُهْلِكَ الْحَرْثَ وَالنَّسْلَ ۚ وَاللهُ لَا يُحِبُّ الْفَسَادَ

"Dan apabila dia berpaling (dari engkau), dia berusaha untuk berbuat kerusakan di bumi, serta merusak tanam-tanaman dan ternak, sedang Allah tidak menyukai kerusakan", (Q.S. Al-Baqarah:205)

Tafsir Ibnu Katsir menekankan bahwa ayat ke-205 surah Al-Baqarah merupakan peringatan terhadap orang-orang yang melakukan kerusakan di bumi dan Allah tidak menyukai kerusakan dan tindakan merusak yang dilakukan oleh orang-orang munafik. Salah satu upaya untuk mengurangi kerusakan di bumi adalah dengan mengurangi aktivitas merokok, karena dengan mengurangi bahkan menghentikan aktivitas tersebut dapat menurunkan polusi yang ditimbulkannya.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan skenario pengujian yang telah dilakukan pada penelitian ini yang melibatkan dua model, yaitu Model A dengan data *train* 80% dan data *test* 20%, dan Model B dengan data *latih* 70% dan data *test* 30%. Setiap model mengalami dua pengujian yang berbeda yaitu pengujian tanpa normalisasi dan mengggunakan normalisasi, kemudian ditambahkan pula dengan pengujian menggunakan *SMOTE* dan seleksi fitur serta tanpa keduanya. Hasil pengujian tersebut dapat dilihat pada tabel (4.41) dan (4.42).

Hasil terbaik dari 16 pengujian yang telah dilakukan terdapat dalam pengujian dengan Model A dan B dengan normalisasi dan *SMOTE* serta Seleksi Fitur. Dengan Model A dengan perbandingan data *train* 80% dan data *test* 20% yang mengimplementasikan normalisasi dan *SMOTE* serta seleksi fitur menghasilkan nilai *Accuracy* 82,50%, *precision* 81,58%, *recall* 89,39% dan *f1-score* 85,31% dan Model B dengan perbandingan data *train* 70% dan data *test* 30% menghasilkan nilai *Accuracy* 81,65%, *precision* 80,73%, *recall* 89,06%, dan *f1-score* 84,69%. Hal tersebut terjadi karena model yang memiliki data yang seimbang dapat terhindar dari hasil *Accuracy* metrik yang kurang akurat, serta proses seleksi fitur yang digunakan bisa menjadikan model tersebut lebih *interpretable* dan lebih cepat dalam melakukan prediksi dibandingkan dengan model yang menggunakan semua fitur. Sehingga dapat disimpulkan bahwa penelitian yang dilakukan dengan

menggunakan algoritma *Random Forest* pada dataset *body signal of smoking* mendapatkan hasil yang baik dalam melakukan klasifikasi terhadap perokok.

5.2 Saran

Peneliti menyadari terdapat beberapa kekurangan dalam penelitian ini dan perlunya kritik maupun saran yang membangun untuk meningkatkan penelitian berikutnya. Perbaikan yang disa*Rank*an tersebut diantaranya:

1. Menggunakan algoritma klasifikasi yang berbeda sehingga dapat membandingkan nilai akurasi, precision, recall dan f1-score yang didapatkan.
2. Dapat menggunakan skenario pengujian yang lebih kompleks untuk dapat meningkatkan hasil akurasi, precision, recall dan f1-score.

DAFTAR PUSTAKA

Adian, I. K., Agung, I. G., & Arya, G. (2023). *Implementasi Random Forest pada Klasifikasi Penyakit Kardiovaskular dengan Hyperparameter Tuning Grid Search*. *2*(November), 219–226.

Aji, A., Leni, M., & Sayed, A. (2015). Isolasi Nikotin Dari Puntung Rokok Sebagai Insektisida. *Jurnal Teknologi Kimia Unimal*, *4*(1), 100–120. http://ft.unimal.ac.id/teknik_kimia/jurnal

Arifin, D. T., & Yunasri, M. A. (2021). PENGARUH PEROKOK AKTIF DIATAS 5 TAHUN TERHADAP KADAR HEMOGLOBIN DI KECAMATAN PANYILEUKAN KOTA BANDUNG. *Jurnal Inovasi Penelitian*, *2*(5), 1655–1660.

Azizah, K. N., Nhita, F., & Kurniawan, I. (2023). *Model Klasifikasi Berbasis Ekspresi Gen Non-Small Cell Lung Carcinoma (NSCLC) pada Wanita Bukan Perokok Menggunakan Metode Ensemble*. *1*(1), 1–7.

Bonaccorso, G. (2017). *Machine Learning Algorithms*. Packt Publishing Ltd.

Breiman, L. E. O. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Ekayanti, I. G. A. S. (2019). ANALISIS KADAR KOLESTEROL TOTAL DALAM DARAH PASIEN DENGAN DIAGNOSIS PENYAKIT KARDIOVASKULER. *International Journal of Applied Chemistry Research*, *1*(1), 6–11.

Esmally, H., Tayefi, M., Doosti, H., Ghayour-Mobarhan, M., Nezami, H., & Amirabadizadeh, A. (2018). *A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes*. *18*(2).

Han, J., Pei, J., & Tong, H. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Hidayat, Sunyoto, A., & Al Fatta, H. (2023). Klasifikasi Penyakit Jantung Menggunakan Decision Tree dan Random Forest. *Jurnal Sistem Komputer Dan Kecerdasan Buatan*, *VII*(September), 31–40.

Issabakhsh, M., Sánchez-Romero, L. M., Le, T. T. T., Liber, A. C., Tan, J., Li, Y., Meza, R., Mendez, D., & Levy, D. T. (2023). Machine learning application for predicting smoking cessation among US adults: An analysis of waves 1-3 of the PATH study. *PLoS ONE*, *18*(6 JUNE), 1–16. https://doi.org/10.1371/journal.pone.0286883

Khatri, P., Neupane, A., Sapkota, S. R., Bashyal, B., Sharma, D., Chhetri, A., Chirag, K. C., Banjade, A., Sapkota, P., & Bhandari, S. (2021). *Case Report Strenuous Exercise-Induced Tremendously Elevated Transaminases Levels in a Healthy Adult : A Diagnostic Dilemma*. *2021*, 5–7.

Kholidha, A. N., Alifariki, L. O., Kedokteran, F., Halu, U., Tenggara, S., & Selabangga, D. (2019). Hubungan kadar high density lipoprotein (hdl) dengan kejadian hipertensi. *Jurnal Profesi Medika : Jurnal Kedokteran Dan Kesehatan*, *13*(2), 74–81.

Kusumasari, P. (2015). *HUBUNGAN ANTARA MEROKOK DENGAN KADAR KOLESTEROL TOTAL PADA PEGAWAI PABRIK GULA TASIKMADU*

KARANGANYAR NASKAH.

Malaeny, C. S., Katuuk, M., & Onibala, F. (2017). HUBUNGAN RIWAYAT LAMA MEROKOK DAN KADAR KOLESTEROL TOTAL DENGAN KEJADIAN PENYAKIT JANTUNG KORONER DI POLIKLINIK JANTUNG RSU PANCARAN KASIH GMIM MANADO. *E-Journal Keperawatan*, *5*.

Muhamad, I., & Matin, M. (2023). *Hyperparameter Tuning menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware*. *9*(1).

Mutoffar, M. M., & Fadillah, A. (2022). Klasifikasi Kualitas Air Sumur Menggunakan Algoritma Random Forest. *Naratif : Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, *4*(2), 138–146. https://doi.org/10.53580/naratif.v4i2.160

Muzayyaroh, & Suyati. (2018). HUBUNGAN KADAR Hb (HAEMOGLOBIN) DENGAN PRESTASI BELAJAR PADA MAHASISWI PRODI D-III KEBIDANAN FIK UNIPDU JOMBANG. *Jurnal Kesehatan Kusuma Husada*.

Najiyah, F., Masfufah, A., Fahmi, N. F., & Izzati, Y. M. P. (2020). Analisis Kadar HDL pada Perokok aktif dewasa usia 31-35 tahun di RT.04 RW.01 Kelurahan Mlajah Bangkalan. *Jurnal Medical*, *2*, 29–33.

Nakamura, M., Yamamoto, Y., Imaoka, W., Kuroshima, T., Toragai, R., & Ito, Y. (2021). *Relationships between Smoking Status , Cardiovascular Risk Factors , and Lipoproteins in a Large Japanese Population*. 942–953.

Nasution, D. P. (2022). *Gambaran Kadar Enzim Aspartat Aminotransferase (Ast) Dan Enzim Alanin Aminotransferase (Alt) Pada Pasien Penderita Sirosis Hati Di Rumah Sakit Efarina Etaham Berastagi*. *1*(5), 992–996.

Qaradhawi, S. M. Y. (1993). *Halal dan Haram dalam Islam*.

Rahmadeyan, A., & Mustakim. (2023). Seleksi Fitur pada Supervised Learning : Klasifikasi Prestasi Belajar Mahasiswa Saat dan Pasca Pandemi COVID-19. *Jurnal Nasional Teknologi Dan Sistem Informasi*, *01*, 21–32.

Song, J., Gao, Y., Yin, P., Li, Y., Li, Y., Zhang, J., Su, Q., Fu, X., & Pi, H. (2021). The random forest model has the best accuracy among the four pressure ulcer prediction models using machine learning algorithms. *Risk Management and Healthcare Policy*, *14*, 1175–1187. https://doi.org/10.2147/RMHP.S297838

Subarkah, P., Risma, W., & Aditya, R. (2022). *Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification*. *14*(2), 120–125.

Syafrina, A. E. (2018). *ANCAMAN PRIVASI DALAM BIG DATA Tentu saja tantangan dalam Big Data*. 138–149.

Teguh, A., Almais, W., Crysdian, C., Fahmi, K., Holle, H., & Roihan, A. (2022). *Smart Assessment Menggunakan Backpropagation Neural Network Smart Assessment using Backpropagation Neural Network*. *21*(3). https://doi.org/10.30812/matrik.v21i3.1382

Ulandhary, Naim, N., Hasan, Z. A., & Armah, Z. (2020). Kadar Hemoglobin, Hitung Jumlah Eritrosit dan nilai hematokrit pada pekerja parkiran basement di kota makassar. *Jurnal Media Analis Kesehatan*, *11*(2), 89–95.

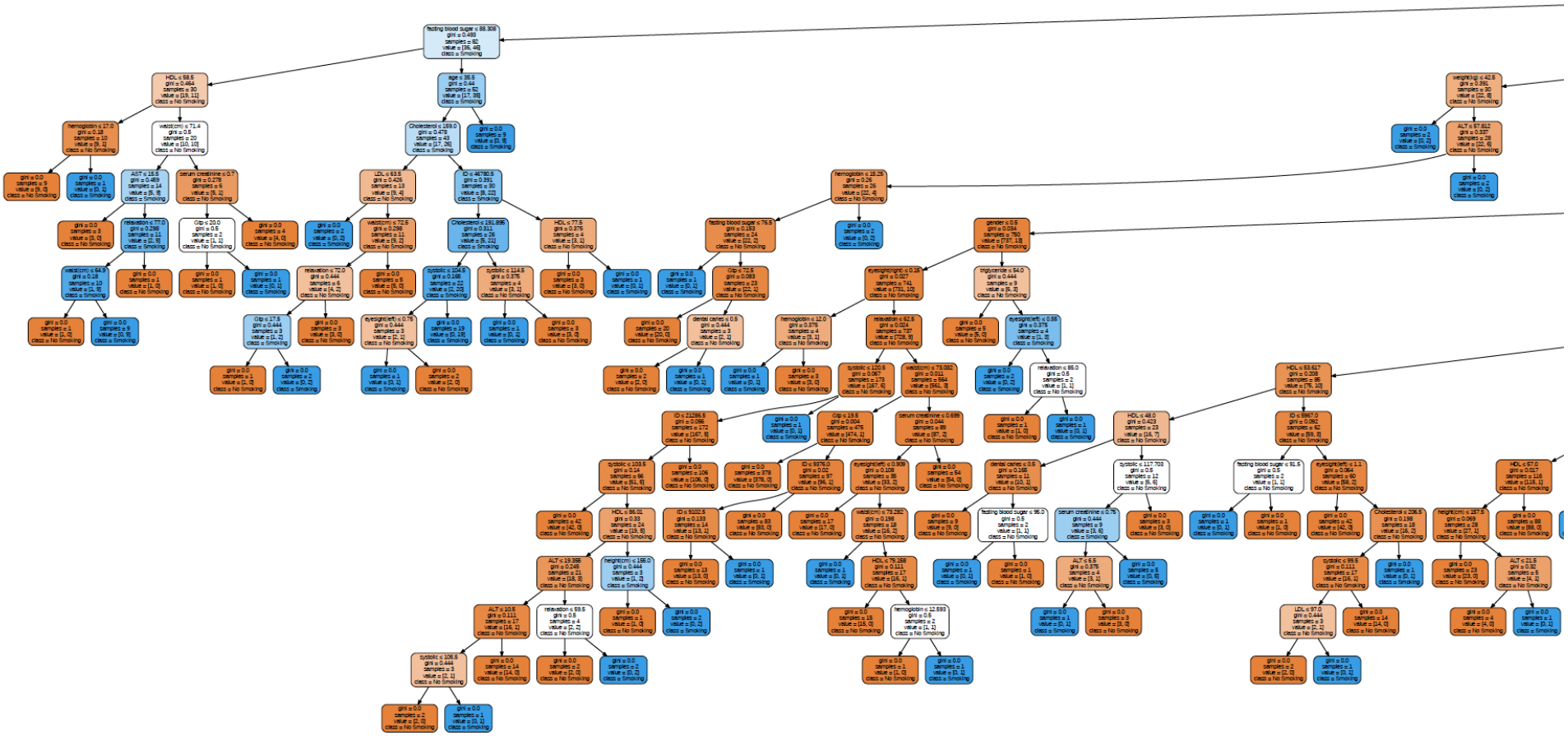
Yazid, R. M., Umbara, F. R., & Sabrina, P. N. (2022). *Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia*. *2*, 46–52.

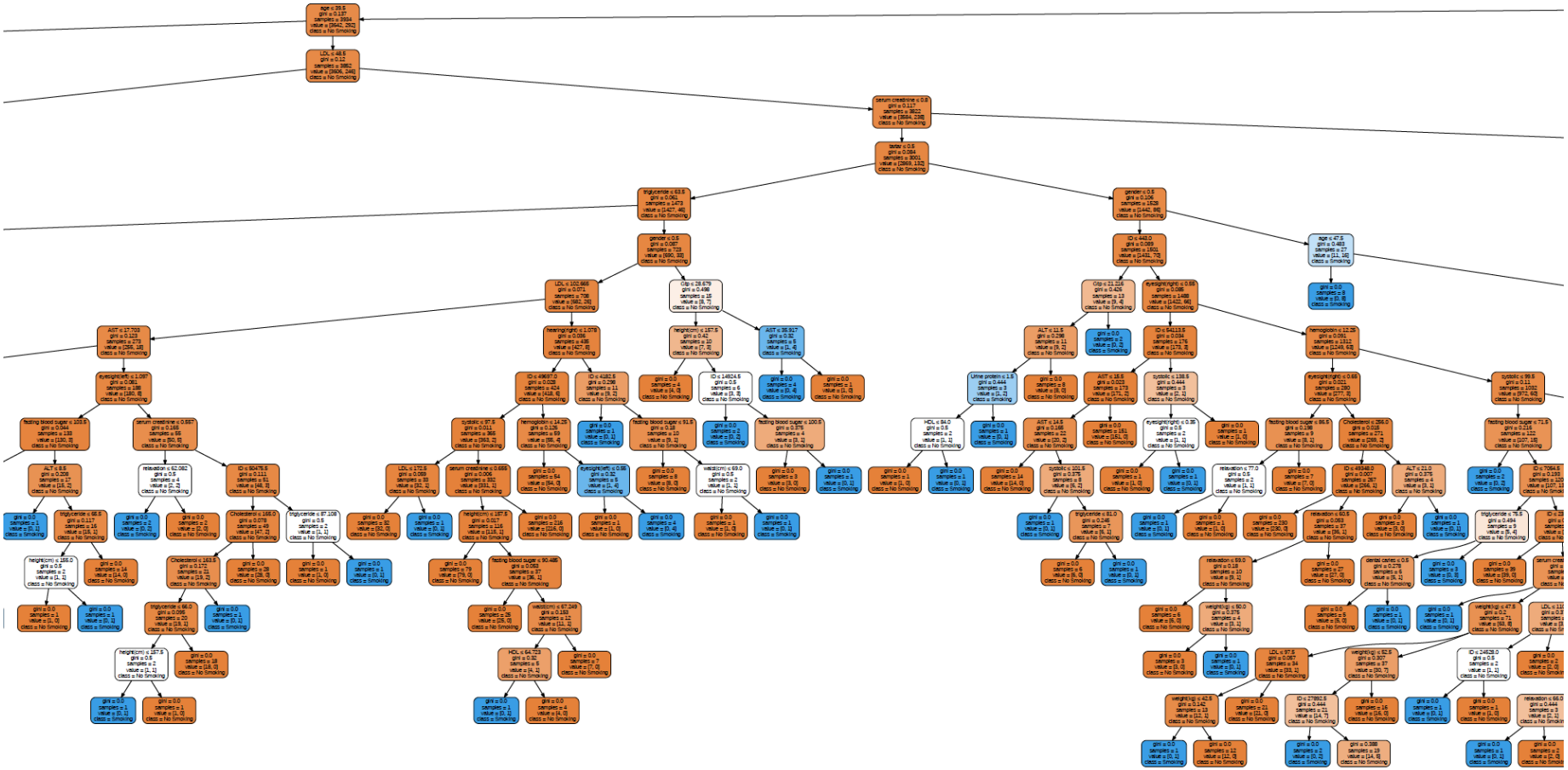
Yu, C. H. (2015). *Exploratory data analysis in the context of data mining and resampling Exploratory data analysis in the context of data mining and resampling . April*. https://doi.org/10.21500/20112084.819

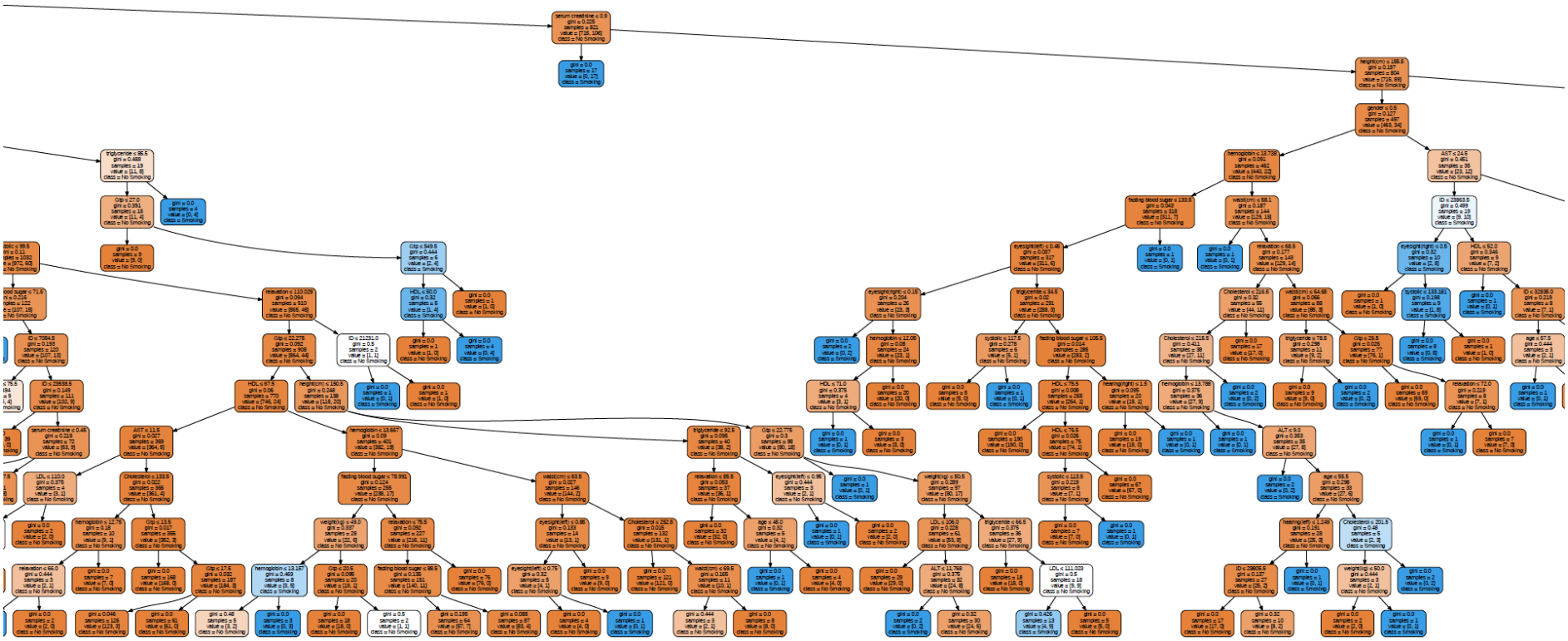
Zhou, Z.-H. (2021). *Machine Learning*. Springer Nature. https://doi.org/https://doi.org/10.1007/978-981-15-1967-3

LAMPIRAN

Lampiran 1. Contoh salah satu *Tree* hasil dari klasifikasi menggunakan *Random Forest*




age ≤ 39.5
gini = 0.137
samples = 3934
value = [3642, 292]
class = No Smoking
LDL ≤ 48.5
gini = 0.13
samples = 3862
value = [3606, 256]
class = No Smoking
serum creatinine ≤ 0.9
gini = 0.112
samples = 3622
value = [3394, 228]
class = No Smoking
tartar ≤ 0.5
gini = 0.084
samples = 3001
value = [2869, 132]
class = No Smoking
triglyceride ≤ 63.5
gini = 0.061
samples = 1473
value = [1427, 46]
class = No Smoking
gender ≤ 0.5
gini = 0.057
samples = 723
value = [690, 33]
class = No Smoking
LDL ≤ 100.666
gini = 0.071
samples = 709
value = [662, 33]
class = No Smoking
AST ≤ 17.709
gini = 0.101
samples = 379
value = [356, 18]
class = No Smoking
eyesight(left) ≤ 1.097
gini = 0.061
samples = 188
value = [180, 8]
class = No Smoking
fasting blood sugar ≤ 103.5
gini = 0.042
samples = 139
value = [136, 3]
class = No Smoking
ALT ≤ 6.5
gini = 0.208
samples = 17
value = [15, 2]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
triglyceride ≤ 66.5
gini = 0.117
samples = 16
value = [15, 1]
class = No Smoking
height(cm) ≤ 159.0
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 14
value = [14, 0]
class = No Smoking
serum creatinine ≤ 0.887
gini = 0.165
samples = 49
value = [45, 4]
class = No Smoking
relaxation ≤ 82.062
gini = 0.5
samples = 4
value = [2, 2]
class = No Smoking
gini = 0.0
samples = 2
value = [2, 0]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
ID ≤ 80478.5
gini = 0.111
samples = 45
value = [43, 2]
class = No Smoking
Cholesterol ≤ 169.0
gini = 0.078
samples = 49
value = [47, 2]
class = No Smoking
Cholesterol ≤ 162.5
gini = 0.172
samples = 21
value = [19, 2]
class = No Smoking
triglyceride ≤ 66.0
gini = 0.095
samples = 20
value = [19, 1]
class = No Smoking
height(cm) ≤ 157.5
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 18
value = [18, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 28
value = [28, 0]
class = No Smoking
triglyceride = 87.108
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
hearing(right) ≤ 1.078
gini = 0.036
samples = 436
value = [427, 9]
class = No Smoking
ID ≤ 49697.0
gini = 0.028
samples = 424
value = [418, 6]
class = No Smoking
systolic ≤ 97.5
gini = 0.011
samples = 360
value = [359, 2]
class = No Smoking
LDL ≤ 172.5
gini = 0.059
samples = 66
value = [62, 1]
class = No Smoking
gini = 0.0
samples = 62
value = [62, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
hemoglobin ≤ 14.25
gini = 0.026
samples = 294
value = [291, 3]
class = No Smoking
serum creatinine ≤ 0.655
gini = 0.006
samples = 282
value = [281, 1]
class = No Smoking
height(cm) ≤ 167.5
gini = 0.017
samples = 116
value = [115, 1]
class = No Smoking
gini = 0.0
samples = 79
value = [79, 0]
class = No Smoking
fasting blood sugar ≤ 90.486
gini = 0.053
samples = 37
value = [36, 1]
class = No Smoking
gini = 0.0
samples = 28
value = [28, 0]
class = No Smoking
waist(cm) ≤ 67.249
gini = 0.188
samples = 9
value = [8, 1]
class = No Smoking
HDL ≤ 64.723
gini = 0.32
samples = 5
value = [4, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 4
value = [4, 0]
class = No Smoking
gini = 0.0
samples = 216
value = [216, 0]
class = No Smoking
eyesight(left) ≤ 0.98
gini = 0.32
samples = 5
value = [1, 4]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 4
value = [0, 4]
class = Smoking
ID ≤ 41828.5
gini = 0.236
samples = 11
value = [9, 2]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
fasting blood sugar ≤ 91.6
gini = 0.18
samples = 10
value = [9, 1]
class = No Smoking
gini = 0.0
samples = 8
value = [8, 0]
class = No Smoking
waist(cm) ≤ 69.0
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
Cig ≤ 29.679
gini = 0.496
samples = 15
value = [8, 7]
class = No Smoking
height(cm) ≤ 157.5
gini = 0.42
samples = 10
value = [7, 3]
class = No Smoking
gini = 0.0
samples = 4
value = [4, 0]
class = No Smoking
ID ≤ 14824.5
gini = 0.5
samples = 6
value = [3, 3]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
fasting blood sugar ≤ 100.9
gini = 0.375
samples = 4
value = [3, 1]
class = No Smoking
gini = 0.0
samples = 3
value = [3, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
AST ≤ 30.917
gini = 0.32
samples = 5
value = [1, 4]
class = Smoking
gini = 0.0
samples = 4
value = [0, 4]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gender ≤ 0.5
gini = 0.105
samples = 1538
value = [1442, 96]
class = No Smoking
ID ≤ 448.0
gini = 0.089
samples = 1491
value = [1431, 70]
class = No Smoking
Cig ≤ 21.316
gini = 0.426
samples = 13
value = [9, 4]
class = No Smoking
ALT ≤ 11.5
gini = 0.298
samples = 11
value = [9, 2]
class = No Smoking
Urine protein ≤ 1.5
gini = 0.444
samples = 3
value = [1, 2]
class = Smoking
HDL ≤ 84.0
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 8
value = [8, 0]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
eyesight(right) ≤ 0.98
gini = 0.085
samples = 1478
value = [1422, 56]
class = No Smoking
AST ≤ 18.5
gini = 0.033
samples = 173
value = [171, 2]
class = No Smoking
AST ≤ 14.5
gini = 0.496
samples = 13
value = [6, 7]
class = Smoking
gini = 0.0
samples = 6
value = [6, 0]
class = No Smoking
gini = 0.0
samples = 161
value = [161, 0]
class = No Smoking
systolic ≤ 138.5
gini = 0.444
samples = 3
value = [1, 2]
class = Smoking
Eyesight(right) ≤ 0.98
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
ID ≤ 641115.5
gini = 0.034
samples = 1305
value = [1278, 27]
class = No Smoking
systolic ≤ 101.5
gini = 0.375
samples = 4
value = [3, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 3
value = [3, 0]
class = No Smoking
relaxation ≤ 77.0
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1301
value = [1278, 23]
class = No Smoking
age ≤ 47.5
gini = 0.483
samples = 47
value = [11, 36]
class = Smoking
gini = 0.0
samples = 8
value = [0, 8]
class = Smoking
hemoglobin ≤ 12.35
gini = 0.091
samples = 1311
value = [1245, 66]
class = No Smoking
eyesight(right) ≤ 0.69
gini = 0.021
samples = 290
value = [277, 6]
class = No Smoking
fasting blood sugar ≤ 86.5
gini = 0.245
samples = 19
value = [16, 3]
class = No Smoking
gini = 0.0
samples = 7
value = [7, 0]
class = No Smoking
Cholesterol ≤ 286.0
gini = 0.02
samples = 271
value = [265, 6]
class = No Smoking
ID ≤ 40948.0
gini = 0.007
samples = 267
value = [266, 1]
class = No Smoking
relaxation ≤ 60.5
gini = 0.083
samples = 37
value = [35, 2]
class = No Smoking
gini = 0.0
samples = 27
value = [27, 0]
class = No Smoking
dental caries ≤ 0.5
gini = 0.278
samples = 10
value = [8, 2]
class = No Smoking
gini = 0.0
samples = 8
value = [8, 0]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
ALT ≤ 21.0
gini = 0.375
samples = 4
value = [3, 1]
class = No Smoking
gini = 0.0
samples = 3
value = [3, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
systolic ≤ 99.5
gini = 0.11
samples = 1021
value = [972, 49]
class = No Smoking
fasting blood sugar ≤ 71.5
gini = 0.426
samples = 13
value = [4, 9]
class = Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
ID ≤ 7064.5
gini = 0.388
samples = 11
value = [4, 7]
class = Smoking
triglyceride ≤ 76.5
gini = 0.494
samples = 9
value = [4, 5]
class = Smoking
gini = 0.0
samples = 99
value = [99, 0]
class = No Smoking
gini = 0.0
samples = 3
value = [0, 3]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
relaxation ≤ 59.0
gini = 0.18
samples = 10
value = [9, 1]
class = No Smoking
gini = 0.0
samples = 6
value = [6, 0]
class = No Smoking
weight(kg) ≤ 60.0
gini = 0.375
samples = 4
value = [3, 1]
class = No Smoking
gini = 0.0
samples = 3
value = [3, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
LDL ≤ 97.5
gini = 0.067
samples = 84
value = [80, 4]
class = No Smoking
weight(kg) ≤ 42.8
gini = 0.142
samples = 13
value = [12, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 12
value = [12, 0]
class = No Smoking
gini = 0.0
samples = 71
value = [71, 0]
class = No Smoking
weight(kg) ≤ 62.5
gini = 0.507
samples = 37
value = [30, 7]
class = No Smoking
ID ≤ 279865.5
gini = 0.444
samples = 21
value = [14, 7]
class = No Smoking
gini = 0.0
samples = 16
value = [16, 0]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
gini = 0.389
samples = 19
value = [14, 5]
class = No Smoking
weight(kg) ≤ 47.9
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
ID ≤ 24528.0
gini = 0.5
samples = 2
value = [1, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
relaxation ≤ 66.0
gini = 0.444
samples = 3
value = [1, 2]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking


serum creatinine ≤ 0.9
gini = 0.225
samples = 823
value = [715, 108]
class = No Smoking
gini = 0.0
samples = 22
value = [0, 27]
class = Smoking
height(cm) ≤ 166.5
gini = 0.197
samples = 804
value = [715, 89]
class = No Smoking
triglyceride ≤ 86.5
gini = 0.488
samples = 19
value = [11, 8]
class = No Smoking
gender ≤ 0.5
gini = 0.157
samples = 497
value = [463, 34]
class = No Smoking
hemoglobin ≤ 13.738
gini = 0.091
samples = 462
value = [440, 22]
class = No Smoking
AST ≤ 24.5
gini = 0.461
samples = 35
value = [23, 13]
class = No Smoking
Gtp ≤ 27.0
gini = 0.391
samples = 16
value = [11, 4]
class = No Smoking
gini = 0.0
samples = 4
value = [0, 4]
class = Smoking
fasting blood sugar ≤ 133.5
gini = 0.043
samples = 318
value = [311, 7]
class = No Smoking
waist(cm) ≤ 58.1
gini = 0.187
samples = 144
value = [129, 15]
class = No Smoking
ID ≤ 23863.5
gini = 0.499
samples = 19
value = [9, 10]
class = Smoking
gini = 0.0
samples = 9
value = [9, 0]
class = No Smoking
Gtp ≤ 549.5
gini = 0.444
samples = 6
value = [2, 4]
class = Smoking
eyesight(left) ≤ 0.46
gini = 0.097
samples = 317
value = [311, 6]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
relaxation ≤ 68.5
gini = 0.177
samples = 143
value = [129, 14]
class = No Smoking
eyesight(right) ≤ 0.5
gini = 0.32
samples = 10
value = [2, 8]
class = Smoking
HDL ≤ 62.0
gini = 0.346
samples = 9
value = [7, 2]
class = No Smoking
HDL ≤ 60.0
gini = 0.32
samples = 5
value = [1, 4]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
eyesight(right) ≤ 0.18
gini = 0.204
samples = 26
value = [23, 3]
class = No Smoking
triglyceride ≤ 34.5
gini = 0.02
samples = 291
value = [286, 5]
class = No Smoking
Cholesterol ≤ 224.5
gini = 0.32
samples = 55
value = [44, 11]
class = No Smoking
waist(cm) ≤ 64.68
gini = 0.066
samples = 88
value = [85, 3]
class = No Smoking
systolic ≤ 130.181
gini = 0.198
samples = 9
value = [1, 8]
class = Smoking
ID ≤ 32838.0
gini = 0.219
samples = 9
value = [7, 2]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
hemoglobin ≤ 12.06
gini = 0.02
samples = 94
value = [93, 1]
class = No Smoking
systolic ≤ 117.5
gini = 0.278
samples = 6
value = [5, 1]
class = No Smoking
fasting blood sugar ≤ 105.5
gini = 0.014
samples = 285
value = [283, 2]
class = No Smoking
Cholesterol ≤ 215.5
gini = 0.411
samples = 33
value = [27, 11]
class = No Smoking
gini = 0.0
samples = 27
value = [27, 0]
class = No Smoking
triglyceride ≤ 78.5
gini = 0.298
samples = 11
value = [9, 2]
class = No Smoking
Gtp ≤ 26.5
gini = 0.026
samples = 77
value = [76, 1]
class = No Smoking
gini = 0.0
samples = 8
value = [8, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
age ≤ 67.5
gini = 0.444
samples = 3
value = [2, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
HDL ≤ 71.0
gini = 0.375
samples = 4
value = [3, 1]
class = No Smoking
gini = 0.0
samples = 20
value = [20, 0]
class = No Smoking
gini = 0.0
samples = 6
value = [6, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
HDL ≤ 79.5
gini = 0.008
samples = 265
value = [264, 1]
class = No Smoking
hearing(right) ≤ 1.5
gini = 0.095
samples = 20
value = [19, 1]
class = No Smoking
hemoglobin ≤ 13.788
gini = 0.375
samples = 4
value = [3, 1]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
gini = 0.0
samples = 9
value = [9, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
relaxation ≤ 72.0
gini = 0.219
samples = 8
value = [7, 1]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 7
value = [7, 0]
class = No Smoking
gini = 0.0
samples = 190
value = [190, 0]
class = No Smoking
HDL ≤ 76.5
gini = 0.026
samples = 75
value = [74, 1]
class = No Smoking
gini = 0.0
samples = 19
value = [19, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
ALT ≤ 9.0
gini = 0.353
samples = 35
value = [27, 8]
class = No Smoking
systolic ≤ 113.5
gini = 0.219
samples = 8
value = [7, 1]
class = No Smoking
gini = 0.0
samples = 67
value = [67, 0]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
age ≤ 65.5
gini = 0.296
samples = 33
value = [27, 6]
class = No Smoking
hearing(left) ≤ 1.249
gini = 0.191
samples = 28
value = [25, 3]
class = No Smoking
Cholesterol ≤ 201.5
gini = 0.48
samples = 5
value = [2, 3]
class = Smoking
ID ≤ 29806.5
gini = 0.137
samples = 27
value = [25, 2]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
weight(kg) ≤ 50.0
gini = 0.444
samples = 3
value = [2, 1]
class = No Smoking
gini = 0.0
samples = 2
value = [0, 2]
class = Smoking
gini = 0.0
samples = 17
value = [17, 0]
class = No Smoking
gini = 0.32
samples = 10
value = [8, 2]
class = No Smoking
gini = 0.0
samples = 2
value = [2, 0]
class = No Smoking
gini = 0.32
samples = 5
value = [1, 4]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 4
value = [0, 4]
class = Smoking
gini = 0.0
samples = 1
value = [1, 0]
class = No Smoking
gini = 0.0
samples = 1
value = [0, 1]
class = Smoking
gini = 0.0
samples = 1
value = [0, 1