TUGAS UAS DATA MINING

"Perbandingan Metode Supervised Learning dalam Mengklasifikasikan Status Merokok Berdasarkan Kondisi Tubuh Seseorang"

Dosen Pengampu : Abu Salam, M. Kom.

Disusun oleh:

Rizka Nugraha (A11.2022.14119)

TEKNIK INFORMATIKA
UNIVERSITAS DIAN NUSWANTORO
2025

I. Pendahuluan

Supervised learning merupakan suatu metode yang terlibat dalam pengembangan kecerdasan buatan, di mana algoritma-algoritma komputer dilatih menggunakan data yang telah diberi label untuk menghasilkan output yang diinginkan. Pendekatan ini bertujuan untuk mengidentifikasi pola-pola dan menemukan korelasi yang terdapat antara input data dan label output yang ada. Supervised learning akan bermanfaat ketika developer memiliki data yang sudah berlabel dengan baik, karena mampu menghasilkan prediksi yang akurat untuk variabel output berdasarkan variabel input yang diberikan.

Supervised learning, yang merupakan salah satu cabang utama dari machine learning, digunakan untuk menyelesaikan berbagai macam masalah dengan bantuan data yang telah dilabeli sebelumnya. Terdapat dua jenis masalah utama yang dapat dipecahkan menggunakan supervised learning, yaitu regresi dan klasifikasi. Pada masalah regresi, variabel output yang diinginkan adalah nilai numerik, sementara pada masalah klasifikasi, tujuannya adalah untuk menentukan kategori atau kelas dari setiap variabel input. Dengan menggunakan metode supervised learning, perusahaan dapat memanfaatkannya untuk menangani berbagai masalah berskala besar. Sebagai contoh, dalam industri e-commerce, supervised learning dapat digunakan untuk menentukan kategori atau label yang sesuai untuk setiap artikel berdasarkan deskripsi atau gambar produk. Selain itu, metode ini juga dapat digunakan untuk melakukan prediksi terkait dengan volume penjualan di masa mendatang berdasarkan data historis penjualan serta faktor-faktor eksternal seperti musim atau tren pasar. Dengan demikian, supervised learning menjadi alat yang sangat berguna bagi perusahaan dalam mengoptimalkan operasinya dan membuat keputusan yang lebih cerdas berdasarkan analisis data yang tepat.

Meskipun supervised learning telah terbukti efektif dalam banyak kasus, terdapat beberapa kelemahan yang perlu diperhatikan. Salah satunya adalah ketergantungan yang tinggi pada data yang telah dilabeli sebelumnya. Hal ini berarti model-model yang dikembangkan melalui supervised learning hanya mampu membuat prediksi yang baik untuk data yang serupa dengan yang telah diberikan pada saat pelatihan. Ketika dihadapkan pada data baru yang berbeda atau memiliki karakteristik yang belum pernah ditemui sebelumnya, kinerja model dapat menurun secara signifikan. Selain itu, supervised learning juga memiliki keterbatasan dalam mengaplikasikan dirinya pada data yang tidak berlabel dalam jumlah besar. Proses pelabelan data

memerlukan waktu, biaya, dan sumber daya manusia yang signifikan, terutama ketika data yang ada sangat besar. Hal ini dapat menjadi hambatan dalam pengembangan model yang memerlukan data yang cukup untuk mendapatkan hasil yang akurat. Masalah lain muncul pada dataset imbalanced adalah ketika terdapat kemungkinan besar bahwa model akan lebih baik dalam mengidentifikasi kelas yang lebih banyak (majority class) daripada kelas yang lebih sedikit (minority class). Ini dapat menyebabkan model tidak mampu mengidentifikasi kelas minority dengan tingkat akurasi yang tinggi, mengurangi kinerja model dalam mengidentifikasi kelas minority, dan mengakibatkan keluaran yang tidak sesuai dengan kebutuhan praktis.

Pada studi kasus ini, akan diaplikasikan beberapa metode supervised learning untuk mengklasifikasikan status merokok seseorang berdasarkan sejumlah tanda yang ditemukan pada tubuh orang tersebut. Latar belakang dalam pemilihan topik ini berasal dari minat yang mendalam dalam bidang kesehatan masyarakat dan keinginan untuk mengungkap wawasan potensial dari data yang ada. Dengan menganalisis sinyal dan kondisi tubuh para perokok, kami bertujuan untuk mendapatkan pemahaman yang lebih mendalam tentang risiko kesehatan yang terkait dengan kebiasaan merokok dan berkontribusi pada pengembangan strategi penghentian merokok yang efektif. Melalui penelitian ini, kami berharap dapat memberikan sumbangan yang berarti dalam meningkatkan kesadaran masyarakat dan membantu individu-individu yang ingin berhenti merokok untuk mencapai tujuan kesehatan mereka.

II. Persiapan Data

2.1 Penjelasan Data

Pada studi ini, dataset yang digunakan merupakan data “Body signal of smoking” yang telah disediakan pada website Kaggle. Dataset ini diunduh dengan format csv yang berasal dari Pemeriksaan Layanan Kesehatan oleh lembaga Asuransi Kesehatan Nasional di Korea Selatan. Pemeriksaan kesehatan ini dilakukan rutin setahun sekali dengan menyasar nasabah Asuransi Kesehatan Nasional yang berusia diatas 40 tahun. Pada penelitian ini, diambil dataset yang merupakan hasil pemeriksaan tahun 2022.

Dataset ini memiliki total 27 kolom, terdiri dari 25 variabel prediktor, 1 variabel target, dan 1 kolom id serta terdiri dari 55.692 records. Variabel target yang digunakan adalah variabel smoking sebagai label klasifikasi. Variabel ini berupa string dengan tipe data nominal yang berisikan “0” jika pasien tersebut tidak merokok dan “1” jika pasien

terindikasi merokok. Berikut adalah variabel prediktor yang digunakan dalam dataset pada kasus ini.

No	Variabel	Keterangan
1	gender	Jenis kelamin (0: Perempuan, 1: Laki-laki)
2	age	Umur (dalam rentang 5 tahunan)
3	height(cm)	Tinggi Badan (dalam rentang 5 cm)
4	weight(kg)	Berat badan (dalam rentang 5 kilo)
5	waist(cm)	Lingkar pinggang
6	eyesight(left)	Penglihatan mata kiri
7	eyesight(right)	Penglihatan mata kanan
8	hearing(left)	Pendengaran telinga kiri (1: Normal, 2: Tidak Normal)
9	hearing(right)	Pendengaran telinga kanan (1: Normal, 2: Tidak Normal)
10	systolic	Tekanan darah sistolik
11	relaxation	Tekanan darah relaksasi
12	fasting blood sugar	Gula darah puasa
13	Cholesterol	Jumlah kolesterol tipe ester dan non ester
14	triglyceride	Jumlah asam lemak
15	HDL	Kolesterol high-density lipoprotein
16	LDL	Kolesterol low-density lipoprotein
17	hemoglobin	Junlah hemoglobin dalam sel darah
18	Urine protein	Jumlah protein dalam urine
19	serum creatinine	Kadar kreatinin dalam darah
20	AST	Nilai uji darah untuk enzim dalam jantung, ginjal, otot

21	ALT	Nilai uji darah untuk enzim dalam sel hati
22	Gtp	Nilai uji darah dalam saluran empedu
23	oral	Uji oral pada pasien
24	dental caries	Keberadaan karies gigi (0: Tidak ada, 1: Ada)
25	tartar	Keberadaan karang gigi (0: Tidak ada, 1: Ada)

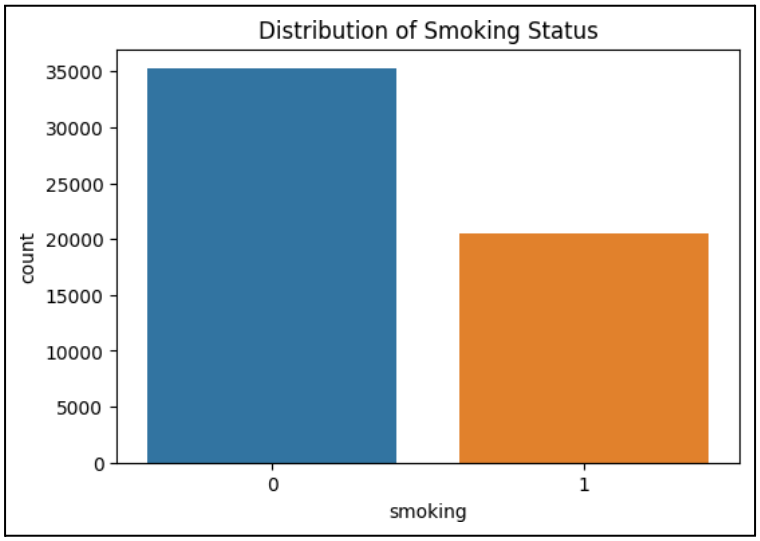
2.2 Preprocessing Data

Berdasarkan pre-processing menggunakan python, didapatkan hasil bahwa pada dataset tidak terdapat missing value untuk setiap kolom atributnya, sehingga tidak perlu dilakukan imputasi. Selanjutnya dilakukan penghapusan kolom “ID” pada dataset. Hal ini dilakukan karena kolom “ID” tidak berpengaruh signifikan dalam proses klasifikasi dan tidak memberikan informasi tambahan dihapus. Kolom “ID” hanya ditujukan untuk penomoran secara urut dan unik pada tiap baris, sehingga penghapusan kolom “ID” dapat dilakukan. Kemudian dilakukan proses encoding data kategorik atau data nominal. Data kategorik perlu dilakukan proses encoding karena sebagian besar algoritma machine learning dan statistik memerlukan input numerik. Sedangkan data kategorikal berupa label atau kategori atau string yang tidak memiliki representasi numerik langsung. Oleh karena itu, perlu dilakukan proses encoding untuk mengubah data kategorik menjadi bentuk numerik. Hasil akhir dari proses encoding membuat seluruh variabel menjadi data dengan tipe data numerik. Encoding dilakukan pada kolom “gender”, “oral”, dan “tartar”.

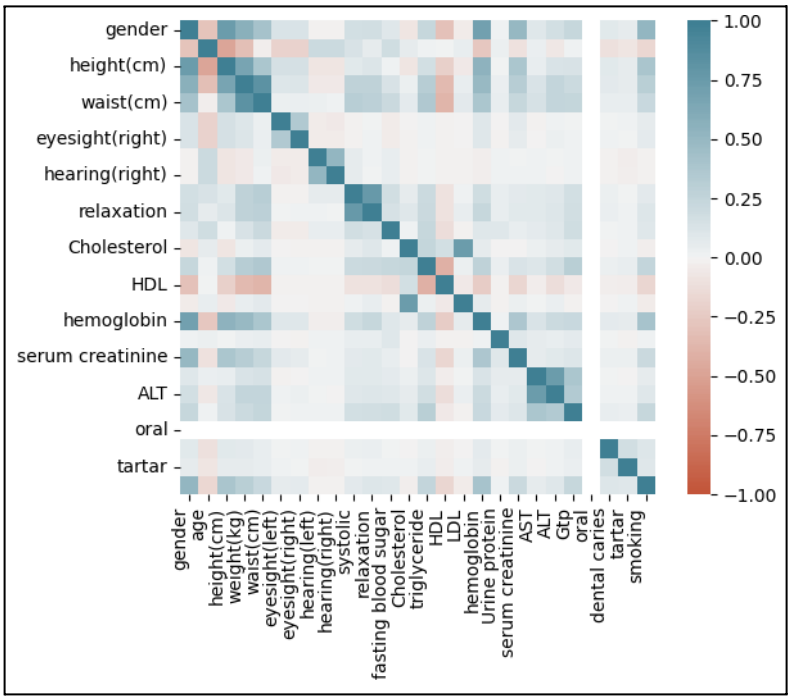
Hasil pada tahap preprocessing menghasilkan data sebanyak 55.692 records data tanpa adanya missing value dengan variabel sebanyak 25 variabel prediktor dan 1 variabel target. Seluruh 22 variabel prediktor telah berubah menjadi tipe data numerik dan nilai yang telah standarisasi pada variabel tersebut. Seluruh variabel ini akan dilakukan klasifikasi dengan menggunakan metode-metode yang telah diajukan.

2.3 Visualisasi Data

Untuk melihat data yang digunakan lebih lanjut, dilakukan visualisasi dari beberapa variabel untuk menentukan langkah selanjutnya yang akan dilakukan. Visualisasi yang dilakukan menggunakan Python dengan bantuan library, seperti seaborn dan matplotlib. Variabel target yang berisikan 0 untuk “tidak merokok” dan 1 untuk “merokok” menunjukkan jumlah data yang cukup berbeda setara dengan “tidak merokok” sekitar 63% total data dan “merokok” sekitar 36% dari total data. Visualisasi ini dapat dilihat pada gambar berikut dalam bentuk diagram batang.



Korelasi antara variabel perlu diperhatikan untuk dapat melakukan beberapa tindakan. Gambar di bawah menunjukkan nilai korelasi antar variabel. Pada dataset ini tidak terdapat variabel yang berkorelasi cukup tinggi sehingga tidak perlu dilakukan tindakan lanjutan untuk mengatasi hal tersebut dan proses analisis bisa dilanjutkan.



2.4 Data Sampling

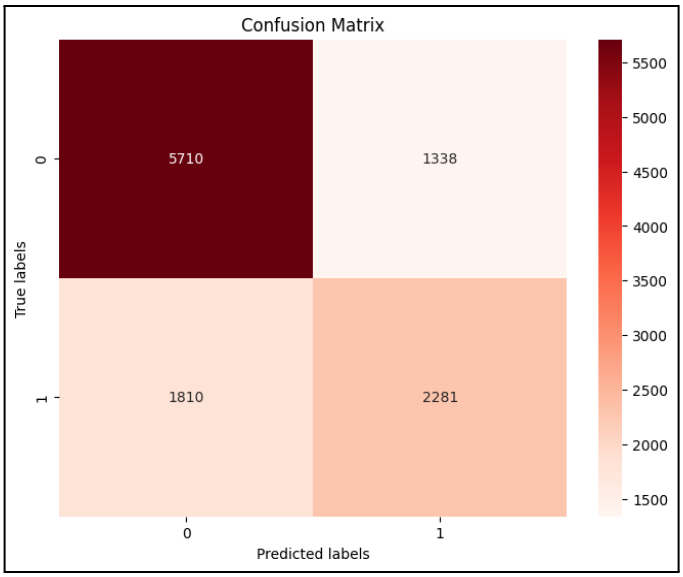
Data hasil tahap preprocessing yang sebanyak 55.692 records digunakan untuk sampling dalam menentukan data training dan data testing. Proses data sampling ini dilakukan dengan menggunakan Python dengan library train_test_split dari sklearn. Data hasil preprocessing sebanyak 80% total data atau sekitar 44.553 digunakan sebagai data training, sisanya sebanyak 20% total data atau sekitar 11.139 digunakan sebagai data testing. Data training akan digunakan sebagai pembangunan model dari metode-metode yang diajukan dan data testing akan digunakan sebagai penilaian metode terbaik dari model-model yang didapat.

III. Metode Klasifikasi

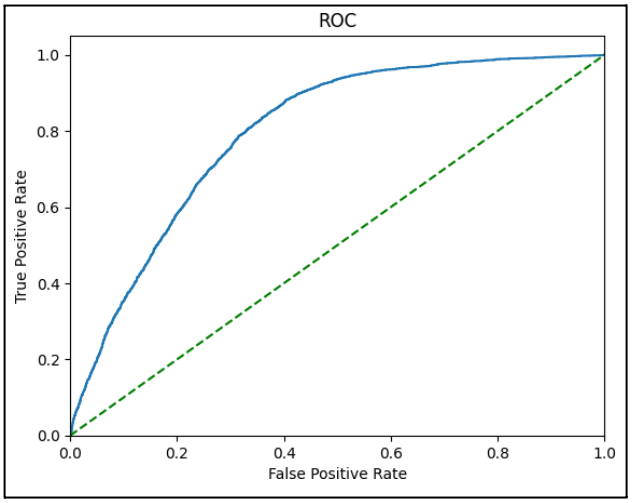
Metode yang digunakan dalam klasifikasi, yaitu Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree, Linear Discrimination Analysis (LDA), dan Naive Bayes. Dalam proses pengaplikasian, data diolah menggunakan Python dengan library sklearn yang digunakan untuk modeling.

3.1 Logistic Regression

Logistic regression adalah metode statistik yang digunakan untuk memprediksi hubungan antara variabel terikat (dependent) dan variabel bebas (independent) yang merupakan jenis analisis regresi yang dilakukan ketika variabel dependen bersifat dikotomis (binary), seperti output yang bisa berupa sukses/gagal, 0/1, benar/salah, atau ya/tidak. Logistic regression bertujuan untuk menghubungkan hubungan antara variabel terikat dan variabel bebas, dan menggunakan satu atau lebih variabel prediktor yang dapat berupa kontinu atau kategorik. Pada kasus ini digunakan variabel terikat berupa seluruh variabel prediktor dan variabel bebasnya adalah variabel target "smoking". Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan logistic regression.



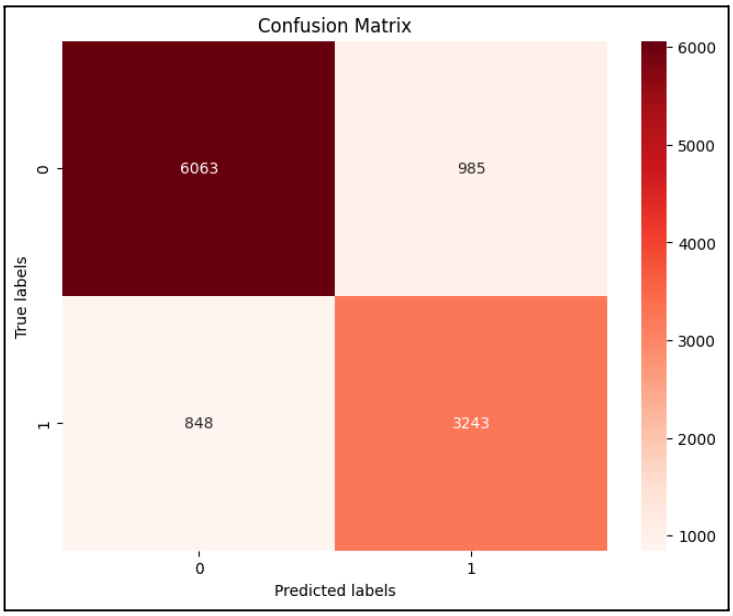
- Sebanyak 5710 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 2281 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 1339 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 1810 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- Logistic regression memiliki tingkat Accuracy sebesar 71,17%, Precision 63,02%, Recall 55,75%



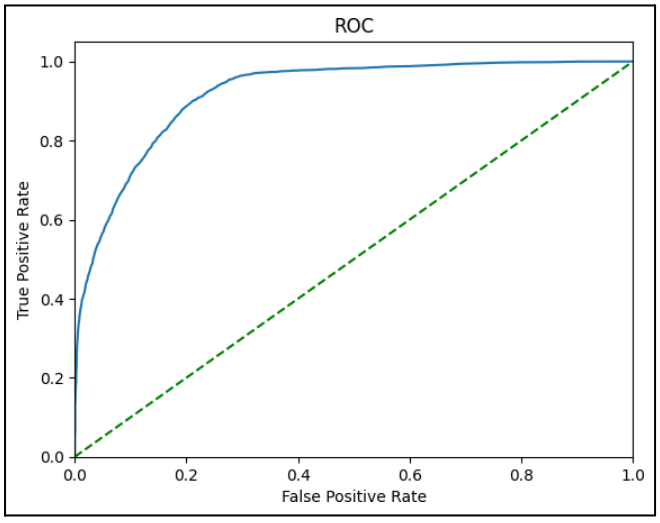
Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat curam, menunjukkan bahwa model dapat meningkatkan TPR dengan cepat tanpa meningkatkan FPR secara signifikan.

3.2 Random Forest

Random forest bekerja dengan cara membangun banyak pohon keputusan (decision trees) secara acak selama proses pelatihan dan menggabungkan hasil prediksi dari pohon-pohon tersebut untuk memperoleh prediksi akhir. Setiap pohon keputusan dibangun dengan menggunakan subset acak dari data pelatihan serta fitur-fitur acak yang dipilih. Dengan memanfaatkan teknik ini, Random Forest mampu mengatasi masalah overfitting yang sering terjadi pada model pohon keputusan tunggal, sehingga menghasilkan model yang lebih stabil dan bias rendah. Pada kasus ini digunakan jumlah estimator sebanyak 200. Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan random forest.

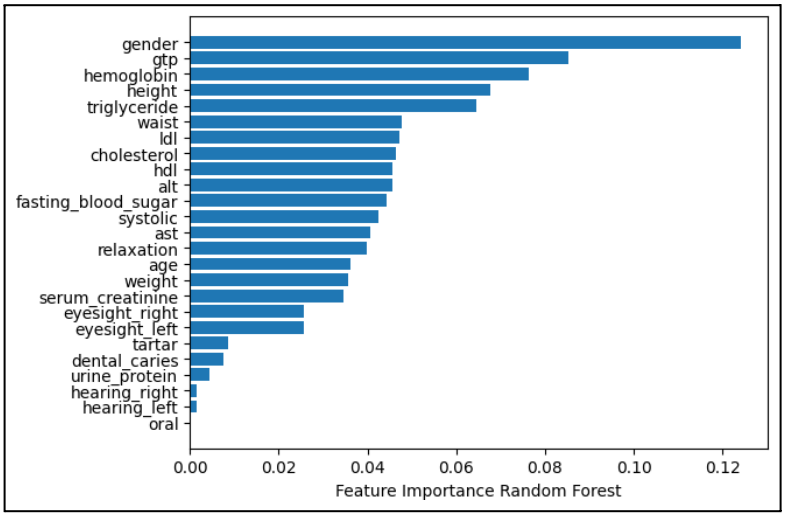


- Sebanyak 6063 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 3243 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 985 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 848 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- Random forest memiliki tingkat Accuracy sebesar 83,54%, Precision 76,70%, Recall 79,27%



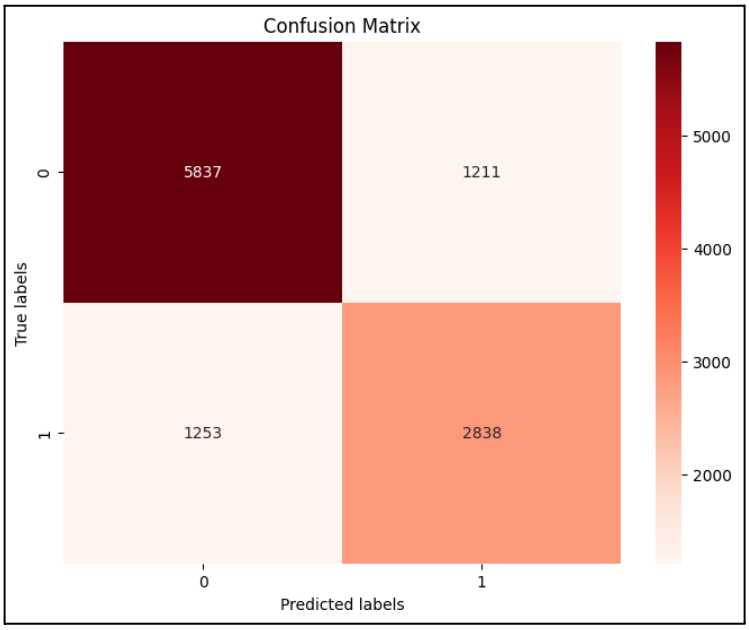
Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat sangat curam, menunjukkan bahwa model dapat meningkatkan TPR dengan cepat tanpa meningkatkan FPR secara signifikan.

Selain itu, random forest memiliki suatu metrik yang digunakan untuk menentukan seberapa penting atau berkontribusinya setiap fitur (variabel) dalam membuat prediksi dengan model Random Forest, yaitu Feature Importance. Pada kasus ini didapatkan beberapa variabel yang dianggap penting melalui algoritma random forest adalah sebagai berikut.

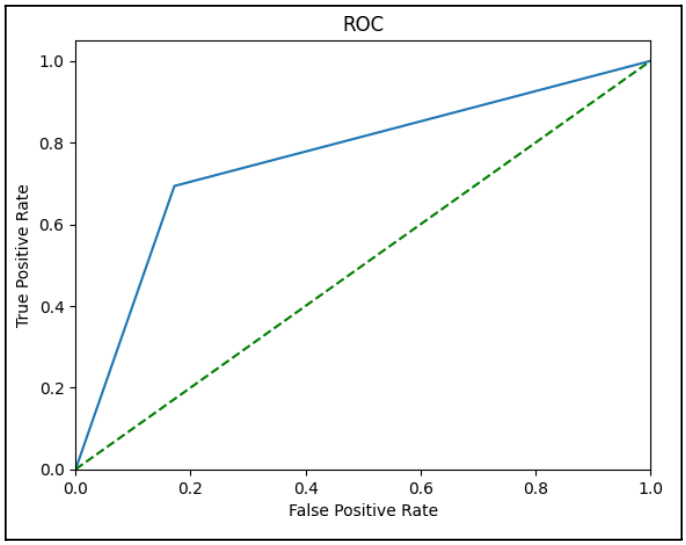


3.3 Decision Tree

Decision Tree bekerja dengan cara mengambil bentuk struktur pohon yang terdiri dari simpul-simpul yang mewakili keputusan, cabang-cabang yang mewakili aturan pengambilan keputusan, dan daun-daun yang mewakili hasil prediksi. Tiap simpul pada pohon keputusan mewakili fitur atau atribut, cabang-cabang mewakili keputusan berdasarkan nilai fitur tersebut, dan daun-daun mewakili hasil prediksi. Proses pembuatan keputusan dimulai dari simpul akar (root node) dan berlanjut ke cabang-cabang sesuai dengan aturan yang ditentukan oleh nilai fitur-fitur. Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan decision tree.

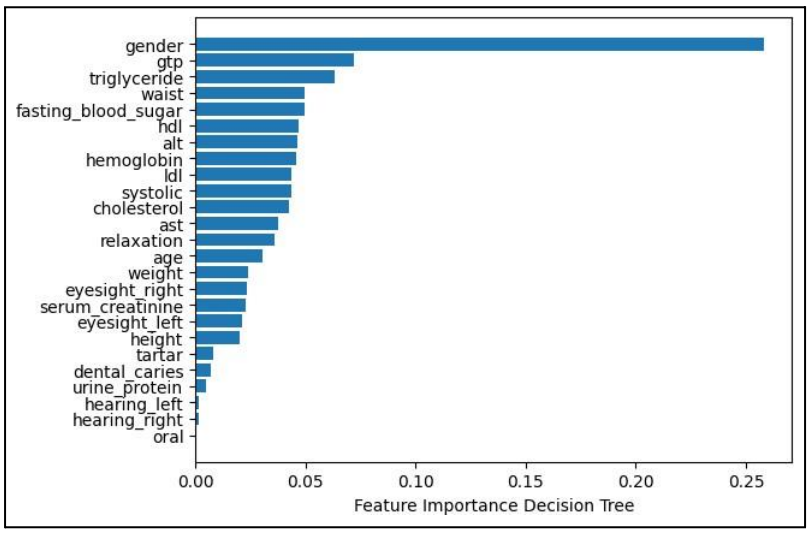


- Sebanyak 5837 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 2838 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 1211 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 1253 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- Decision tree memiliki tingkat Accuracy sebesar 77,87%, Precision 70,09%, Recall 69,37%



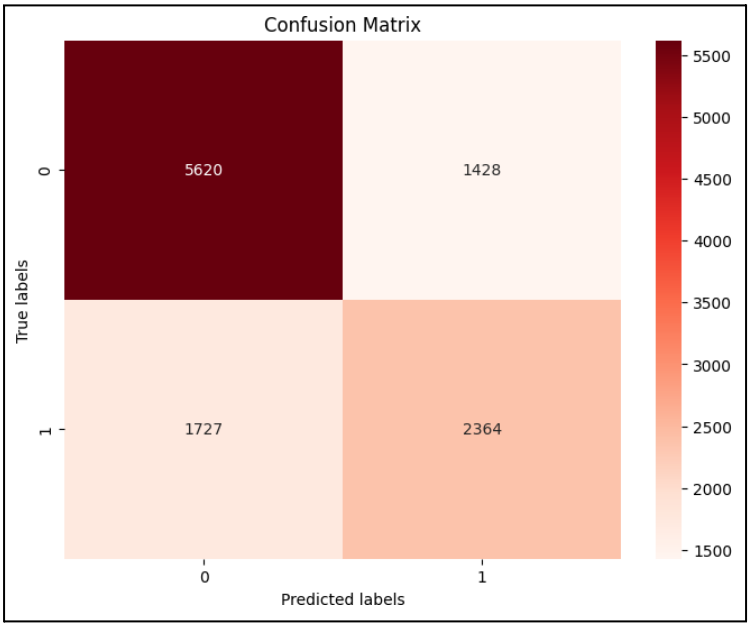
Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat relatif tidak terlalu curam dengan ada patahan saat FPR=0,2, menunjukkan bahwa model dapat meningkatkan TPR dengan relatif baik tanpa meningkatkan FPR secara signifikan.

Selain itu, decision tree memiliki suatu metrik yang digunakan untuk menentukan seberapa penting atau berkontribusinya setiap fitur (variabel) dalam membuat prediksi dengan model Decision Tree , yaitu Feature Importance. Pada kasus ini didapatkan beberapa variabel yang dianggap penting melalui algoritma decision tree adalah sebagai berikut.



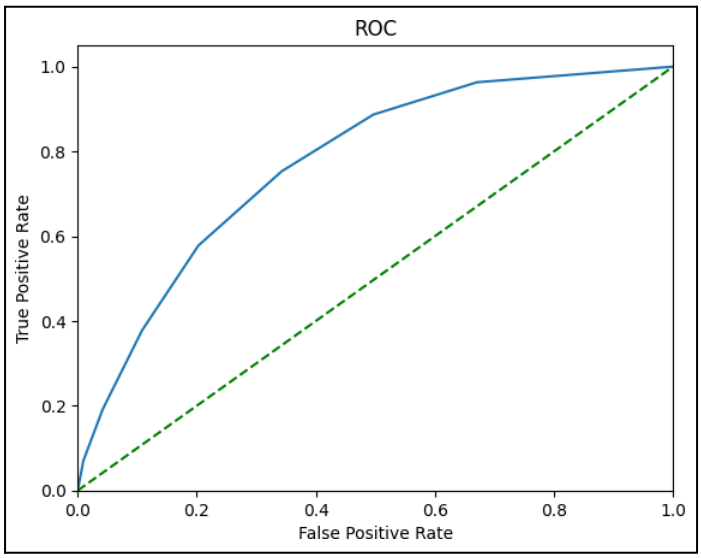
3.4 KNN

K-Nearest Neighbors (KNN) adalah metode klasifikasi yang mencari sekelompok k-objek dalam suatu set training data yang paling dekat dengan objek uji, dan berdasarkan dominasi kelas tertentu di environment tersebut dan menetapkan labelnya. Algoritma ini melibatkan beberapa elemen kunci, yaitu: (i) himpunan objek berlabel yang digunakan untuk mengevaluasi kelas objek uji, (ii) metrik jarak atau kesamaan yang digunakan untuk menghitung kedekatan, (iii) nilai k, jumlah neighbors (tetangga) terdekat, dan (iv) metode yang digunakan untuk menentukan kelas objek target berdasarkan kelas dan jarak k neighbors (tetangga) terdekat. Pada kasus ini digunakan jenis ukuran jarak Euclidean untuk menghitung jarak antara dua titik dengan jumlah neighbors sebanyak 7. Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan KNN.



- Sebanyak 5620 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 2364 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 1428 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 1727 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.

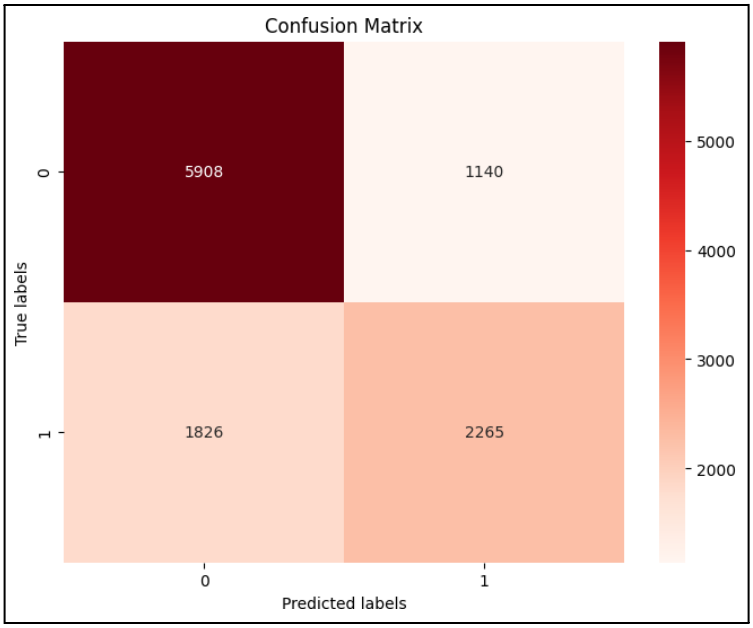
- KNN memiliki tingkat Accuracy sebesar 71,67%, Precision 62,34%, Recall 57,78%



Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat relatif tidak terlalu curam bahkan cenderung mendekati garis diagonal, menunjukkan bahwa model dapat meningkatkan TPR dengan relatif baik tapi tidak sebaik model lain,

3.5 SVM

SVM (Support Vector Machine) adalah algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. Dalam konteks klasifikasi, SVM mencoba untuk membagi data ke dalam dua kelas dengan cara menemukan hyperplane terbaik yang memisahkan kedua kelas secara optimal dalam ruang fitur. Hyperplane ini dipilih sedemikian rupa sehingga jarak antara hyperplane dan titik-titik terdekat (support vectors) dari kedua kelas adalah maksimum. SVM sangat efektif dalam menangani data dengan dimensi tinggi dan dataset yang kompleks, yang mana sejalan dengan dataset yang digunakan pada kasus ini. SVM juga memiliki kemampuan untuk menangani overfitting dengan baik. Pada kasus ini digunakan kernel rbf dan parameter C sebesar 1. Berikut adalah confusion matrix yang dihasilkan menggunakan algoritma SVM.

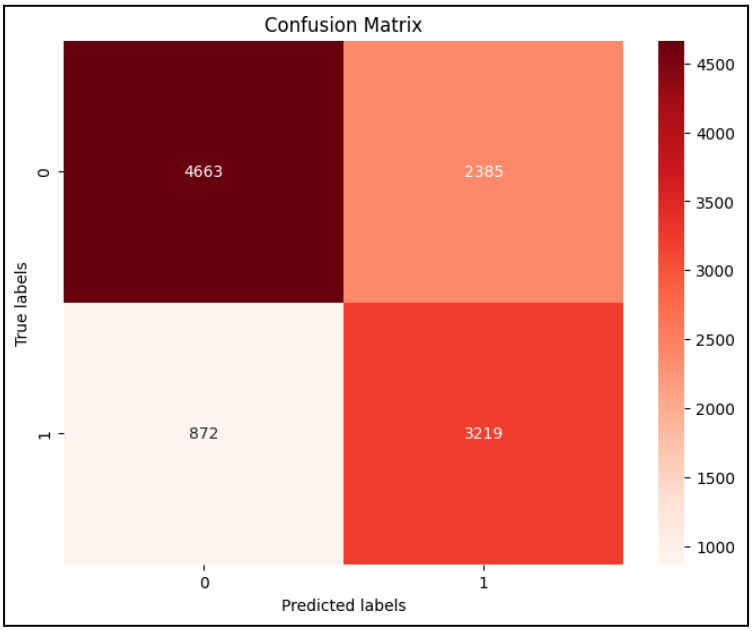


- Sebanyak 5908 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 2265 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 1140 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 1826 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- SVM memiliki tingkat Accuracy sebesar 73,37%, Precision 66,51%, Recall 55,36%

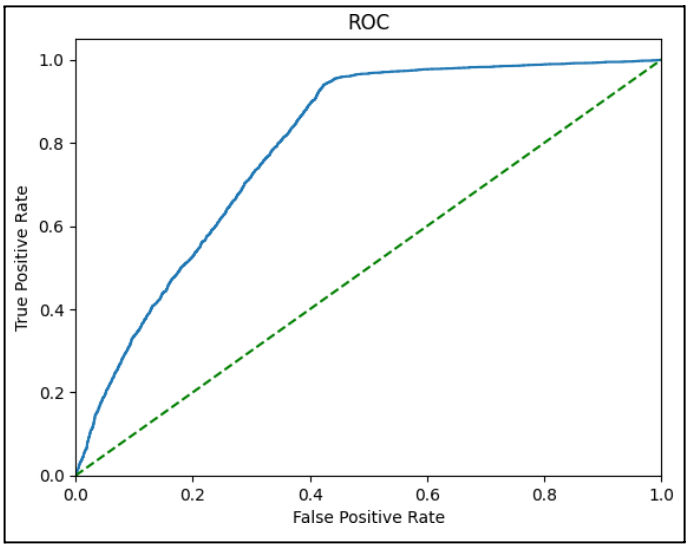
3.6 Naive Bayes

Naive Bayes adalah algoritma klasifikasi yang memprediksi kelas suatu instance/data dengan menghitung probabilitas dari setiap kelas berdasarkan fitur-fitur yang diamati. Algoritma ini bekerja dengan mengasumsikan independensi antara fitur-fitur, sehingga dapat menghitung probabilitas secara efisien. Dengan menggunakan teorema Bayes, Naive Bayes menggabungkan probabilitas prior dan likelihood untuk memperoleh probabilitas posterior, yang digunakan untuk memprediksi kelas. Pada kasus ini, digunakan algoritma Gaussian Naive Bayes yang mana merupakan varian algoritma Naive Bayes yang menggunakan distribusi Gaussian (distribusi

normal) dan variabel kontinu. Model ini digunakan untuk mencari rata-rata dan standar deviasi dari masing-masing kelas. Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan Naive Bayes.



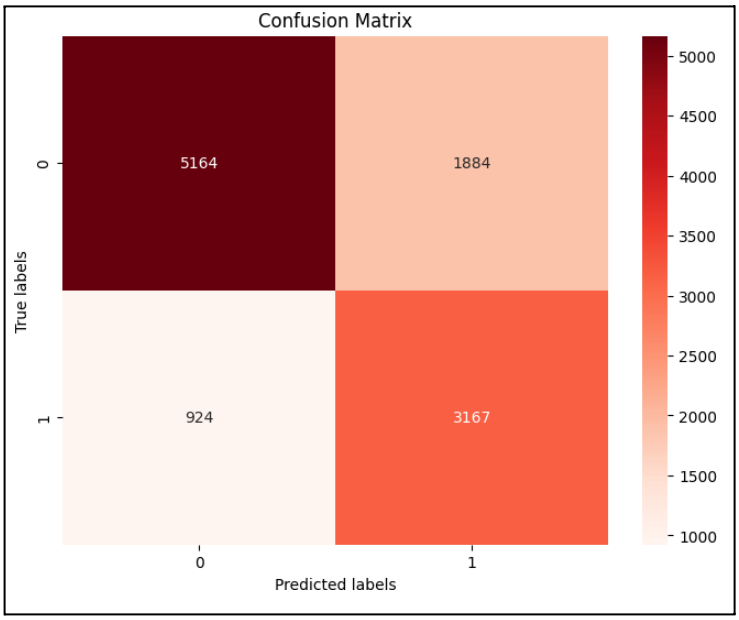
- Sebanyak 4663 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 3219 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 2385 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 872 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- Naive Bayes memiliki tingkat Accuracy sebesar 70,76%, Precision 57,44%, Recall 78,68%



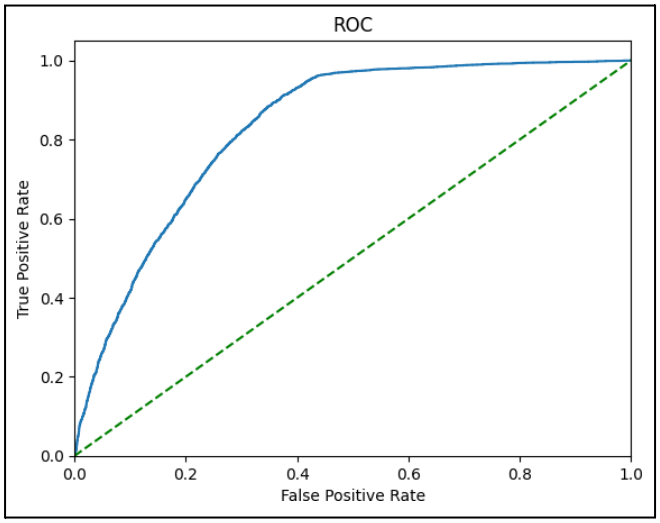
Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat relatif tidak terlalu curam bahkan cenderung mendekati garis diagonal pada awal-awal, menunjukkan bahwa model dapat meningkatkan TPR dengan relatif baik tapi tidak sebaik model lain,

3.7 LDA

Linear Discriminant Analysis (LDA) bekerja dengan cara mengidentifikasi pemisah linear yang optimal antara kelas-kelas dalam data. Prosesnya dimulai dengan menghitung rerata (mean) dari setiap kelas dan matriks dispersi (covariance matrix) dari seluruh dataset. Selanjutnya, LDA menggunakan informasi ini untuk membangun pemisah linear, yang dikenal sebagai linear discriminants, yang digunakan untuk memproyeksikan data ke dimensi yang lebih rendah. Proyeksi dilakukan sedemikian rupa sehingga variabilitas antara kelas maksimum sementara variabilitas dalam setiap kelas minimum. Ketika mendapat data baru, LDA menghitung nilai discriminant function untuk setiap kelas dan memprediksi kelas yang sesuai dengan nilai discriminant tertinggi. Dengan demikian, LDA dapat memisahkan dan mengklasifikasikan data ke dalam kelas yang sesuai berdasarkan pemisah linear yang telah dipelajari dari data training. Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan LDA.



- Sebanyak 5164 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 3167 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 1884 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 924 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- LDA memiliki tingkat Accuracy sebesar 74,79%, Precision 62,70%, Recall 77,41%

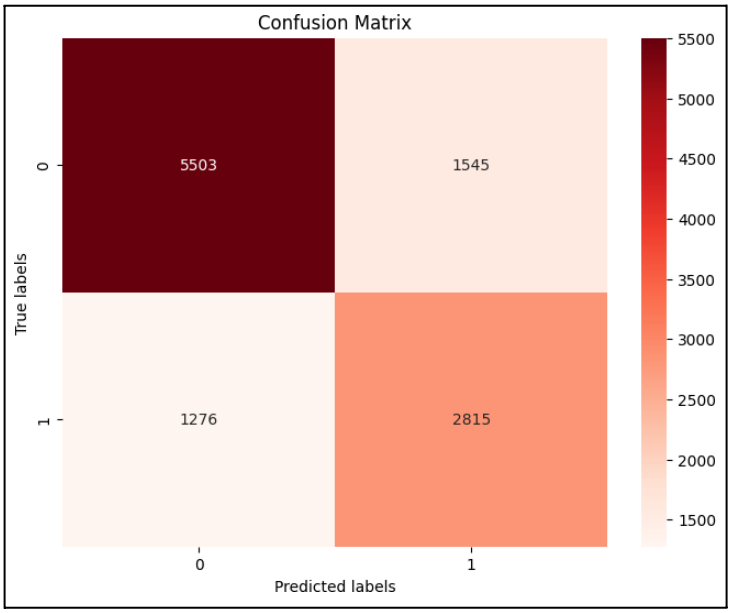


Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat relatif curam menunjukkan bahwa model dapat meningkatkan TPR dengan relatif baik.

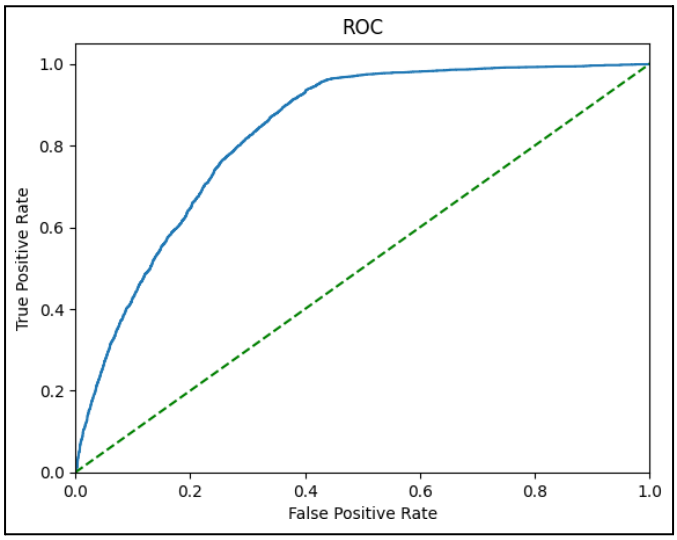
3.8 LASSO

LASSO(Least Absolute Shrinkage and Selection Operator) dalam konteks klasifikasi adalah metode regularisasi yang digunakan untuk mengurangi overfitting dan melakukan seleksi fitur secara otomatis. Lasso menggabungkan fungsi biaya dengan penalti dari nilai absolut koefisien, yang mengakibatkan beberapa koefisien menjadi nol. Dengan demikian, Lasso efektif dalam memilih subset fitur yang paling penting untuk model klasifikasi, sambil mengurangi kompleksitas model. Proses optimasi dalam Lasso melibatkan pencarian nilai koefisien yang meminimalkan fungsi biaya yang dimodifikasi. Ini dapat diselesaikan menggunakan teknik optimasi seperti koordinat turun, gradien turun, atau solusi tertutup tergantung pada konteks dan ukuran data.

Keuntungan utama dari Lasso dalam konteks klasifikasi adalah kemampuannya untuk menghasilkan model yang lebih sederhana dengan mempertahankan hanya fitur-fitur yang paling penting, yang pada gilirannya dapat menghasilkan model yang lebih mudah diinterpretasi dan umumnya memiliki kinerja yang lebih baik pada data uji yang tidak dilihat. Berikut adalah confusion matrix dan kurva ROC dari hasil klasifikasi menggunakan LASSO.



- Sebanyak 5503 pasien yang tidak merokok berhasil diidentifikasi model dengan benar sebagai pasien yang tidak merokok.
- Sebanyak 2815 pasien yang merokok berhasil diidentifikasi model dengan benar sebagai pasien yang merokok.
- Sementara itu ada 1545 pasien yang tidak merokok, tetapi diidentifikasi sebagai pasien yang merokok oleh model dan 1276 pasien yang merokok, tetapi diidentifikasi model sebagai pasien yang tidak merokok.
- LASSO memiliki tingkat Accuracy sebesar 74,67%, Precision 64,56%, Recall 68,80%



Sementara itu, pada Kurva ROC yang terbentuk berada di atas garis diagonal, yang menunjukkan bahwa model memiliki True Positive Rate (TPR) yang lebih tinggi daripada False Positive Rate (FPR) untuk semua nilai threshold. Kurva ROC yang terlihat relatif curam menunjukkan bahwa model dapat meningkatkan TPR dengan relatif baik.

IV. Interpretasi dan Perbandingan Akurasi Model

Setelah dilakukan pengaplikasian beberapa model pada data training (80% total data), tiap model lalu diimplementasikan dengan menggunakan data testing. Evaluasi model dilakukan dengan membandingkan nilai accuracy, precision, dan recall. Pemilihan model terbaik dilihat berdasarkan nilai accuracy tertinggi dan jika terdapat kesamaan nilai dengan model lain dapat mempertimbangkan nilai dari recall dan precision. Berikut ini merupakan hasil dari data testing dengan beberapa model supervised learning yang diajukan.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	71,74	59,17	63,03	55,76
Decision Tree	77,88	69,73	70,09	69,37
Random Forest	**83,54**	**77,97**	**76,70**	**79,27**
KNN	71,68	59,98	62,34	57,79
SVM	73,37	60,43	66,52	55,37
Naïve Bayes	70,76	66,41	57,44	78,68
LDA	74,79	69,28	62,70	77,41
LASSO	74,64	64,56	68,80	66,61

Berdasarkan hasil yang didapatkan, model Random Forest memiliki nilai accuracy sebesar 83,54 yang merupakan nilai accuracy tertinggi dibandingkan dengan model lainnya. Artinya, model Random Forest berhasil melakukan prediksi pada 83,53% dari data yang telah digunakan untuk pemodelan. Nilai precision untuk model Random Forest mendapatkan nilai

sebesar 77,97 yang mana nilai tersebut juga tertinggi dibandingkan dengan model lainnya. Sedangkan untuk nilai recall, model Random Forest memiliki nilai sebesar 76,70 yang mana nilai tersebut juga tertinggi dibandingkan metode lain.

V. Kesimpulan

Berdasarkan hasil studi, dapat disimpulkan bahwa metode Random Forest merupakan metode terbaik untuk mengklasifikasikan status merokok seseorang berdasarkan tanda-tanda kondisi tubuh, dibandingkan dengan metode-metode lainnya seperti K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree, Linear Discriminant Analysis (LDA), dan Naive Bayes, metode Random Forest menunjukkan performa terbaik dengan nilai akurasi tertinggi sebesar 83,54%, presisi sebesar 77,97%, recall sebesar 76,70%, dan F1 Score sebesar 79,27%. Dengan demikian, Random Forest direkomendasikan sebagai pilihan utama bagi instansi kesehatan atau masyarakat yang ingin mengetahui status merokok seseorang dengan hanya melihat kondisi tubuh orang tersebut.

Selain itu, Feature importance pada random forest didominasi oleh variabel “gender”, “gtp”, dan “hemoglobin” sehingga hal-hal tersebut bisa menjadi fokus utama untuk diperhatikan dalam menentukan status merokok seseorang. Selain itu, terdapat beberapa variabel yang berkontribusi utama terhadap pembentukan model diantaranya, “height”, “triglyceride”, dan “waist”.

VI. Lampiran

Source Code pengerjaan tugas dapat diakses melalui tautan berikut: https://github.com/rizkanugrha/DATAMINING-A11.2022.14119-UAS