PENERAPAN ALGORITMA RANDOM FOREST UNTUKKLASIFIKASI STATUS PEROKOK BERDASARKAN DATA FISIOLOGIS

Fajrur Rohman[1], Rudi Kurniawan[2], Bani Nurhakim[3], Indra Wiguna Marthanu[4], Kaslani[5].

Program Studi Teknik Informatika[12]
Program Studi Manajemen Informatika[3]
Program Studi Komputerisasi Akuntansi[45]

STMIK IKMI Cirebon
https://ikmi.ac.id/page/18/?lang=de
fajrur71@gmail.com

(*) Corresponding Author : fajrur71@gmail.com
Published : 30 Mei 2026

Abstract— *This study aims to develop a smoking status classification model using a Random Forest algorithm based on physiological parameters from clinical data. This approach offers a more objective alternative to survey methods or self-reporting, which are prone to bias and underreporting. The data used is an open dataset with a total of 55,692 rows and 27 columns covering various physiological parameters, such as height, hemoglobin, blood pressure, and demographic information. This study consisted of five main stages, namely data acquisition, preprocessing, data division, model training, and evaluation. The preprocessing process included Exploratory Data Analysis (EDA), checking for missing values and duplicate data, outlier detection, categorical variable encoding, and numerical data standardization. The Random Forest baseline model produced an accuracy of 75.65%. After tuning using RandomizedSearchCV, the model performance improved to 76.38%. Feature importance analysis showed that the most influential features for smoking status classification were gender_male, height_cm, and hemoglobin. These results indicate that the Random Forest algorithm with tuning can effectively utilize physiological parameters to classify smoking status and contribute to the development of data-based health screening systems.*

Keywords: *smoking classification, random forest, physiological parameters, machine learning, data preprocessing.*

Abstrak— Penelitian ini bertujuan untuk mengembangkan model klasifikasi status merokok menggunakan algoritma Random Forest berbasis parameter fisiologis dari data klinis. Pendekatan ini menawarkan alternatif yang lebih objektif dibandingkan metode survei atau pelaporan mandiri yang rentan terhadap bias dan *underreporting*. Data yang digunakan merupakan dataset terbuka dengan total 55.692 baris dan 27 kolom yang mencakup berbagai parameter fisiologis, seperti tinggi badan, hemoglobin, tekanan darah, dan informasi demografis. Penelitian ini terdiri atas lima tahapan utama, yaitu akuisisi data, prapemrosesan, pembagian data, pelatihan model, dan evaluasi. Proses prapemrosesan mencakup *Exploratory Data Analysis* (EDA), pengecekan nilai hilang dan data duplikat, deteksi *outlier*, *encoding* variabel kategorikal, dan standarisasi data numerik. Model *baseline* Random Forest menghasilkan akurasi sebesar 75,65%. Setelah dilakukan *tuning* menggunakan RandomizedSearchCV, performa model meningkat menjadi 76,38%. Analisis *feature importance* menunjukkan bahwa fitur paling berpengaruh terhadap klasifikasi status merokok adalah gender_male, height_cm, dan hemoglobin. Hasil ini menunjukkan bahwa algoritma Random Forest dengan *tuning* dapat secara efektif memanfaatkan parameter fisiologis untuk mengklasifikasikan status merokok, serta memberikan kontribusi terhadap pengembangan sistem skrining kesehatan berbasis data.

Keyword: klasifikasi merokok, random forest, parameter fisiologis, *machine learning*, *preprocessing* data.

INTRODUCTION

Merokok masih menjadi salah satu faktor risiko utama berbagai penyakit kronis seperti penyakit kardiovaskular, kanker paru-paru, penyakit paru obstruktif kronis (PPOK), dan gangguan metabolik. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa konsumsi tembakau berkontribusi terhadap jutaan kematian setiap tahun di seluruh dunia. Upaya deteksi status merokok selama ini umumnya dilakukan melalui

kuesioner atau pelaporan mandiri (self-report). Namun, pendekatan ini memiliki kelemahan utama berupa bias pelaporan, underreporting, ketidakakuratan akibat faktor sosial dan psikologis.

Perkembangan pesat di bidang machine learning membuka peluang baru untuk melakukan klasifikasi status merokok secara lebih objektif dan berbasis data fisiologis. Sejumlah penelitian sebelumnya mengeksplorasi penggunaan berbagai algoritma seperti Logistic Regression, Support Vector Machine (SVM), Decision Tree, Neural Network, dan mendeteksi atau Random Forest dalam memprediksi perilaku merokok. Sinha et al. (2025) menunjukkan bahwa Random Forest dan SVM termasuk algoritma yang memiliki performa kuat dalam klasifikasi status merokok berbasis data kesehatan. Ebrahimi et al. (2024) menerapkan pendekatan explainable AI pada data EHR untuk mengklasifikasikan status merokok dengan menekankan interpretabilitas model. Sementara itu, Aishwarya et al. (2025) memanfaatkan dataset klinis publik untuk mengidentifikasi fitur-fitur fisiologis yang memiliki hubungan signifikan dengan status merokok, seperti tekanan darah, indeks massa tubuh, dan kadar hemoglobin.

Meskipun demikian, sebagian besar penelitian terdahulu masih memiliki beberapa keterbatasan. Pertama, banyak studi masih bergantung pada data survei atau kombinasi data subjektif dan objektif, sehingga masih berpotensi mengandung bias. Kedua, tidak semua penelitian berfokus pada pemanfaatan dataset fisiologis skala besar yang sepenuhnya bersifat objektif. Ketiga, aspek interpretabilitas model sering kali belum mendapat perhatian yang memadai, padahal sangat penting dalam konteks kesehatan agar hasil model dapat dipahami dan diterima oleh praktisi medis serta pembuat kebijakan.

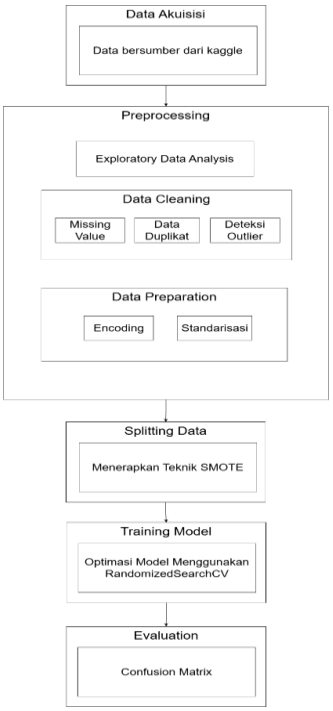
Berdasarkan kondisi tersebut, terdapat kesenjangan penelitian (research gap) yang jelas, yaitu kebutuhan akan model klasifikasi status merokok berbasis parameter fisiologis murni, menggunakan algoritma yang tidak hanya memiliki performa tinggi tetapi juga mampu memberikan interpretasi terhadap faktor-faktor yang paling berpengaruh. Selain itu, masih terbatas penelitian yang memanfaatkan dataset publik berskala besar dengan pendekatan sistematis serta preprocessing evaluasi yang model yang komprehensif.

Oleh karena itu, penelitian ini dilakukan untuk mengisi kesenjangan tersebut dengan mengembangkan model klasifikasi status merokok berbasis parameter fisiologis menggunakan algoritma Random Forest. Penelitian ini juga mengintegrasikan analisis feature telah importance sebagai bentuk interpretabilitas model, sehingga tidak hanya menghasilkan prediksi, tetapi juga memberikan wawasan mengenai parameter fisiologis yang berkontribusi signifikan dalam menentukan status merokok.

MATERIALS AND METHODS

Tahapan Penelitian

Penelitian ini dirancang untuk mengembangkan model klasifikasi status perokok berbasis parameter fisiologis menggunakan algoritma Random Forest. Pendekatan penelitian bersifat kuantitatif dan eksperimental dengan tahapan sistematis mulai dari analisis data, perancangan arsitektur proses, penerapan metode klasifikasi, hingga implementasi dan evaluasi model.



Gambar 1 Tahapan Penelitian

Dengan menggunakan dataset fisiologis berskala besar serta dukungan pustaka machine learning berbasis Python, seluruh proses penelitian diimplementasikan secara komputasional dan terkontrol. Melalui metode ini, diharapkan hasil penelitian tidak hanya menghasilkan model dengan akurasi yang baik, tetapi juga memberikan kontribusi ilmiah dalam pengembangan sistem klasifikasi berbasis data fisiologis yang diinterpretasikan.

Data Akuisisi

Pada tahap ini, data yang digunakan dalam penelitian diperoleh dari platform Kaggle, yaitu dataset publik yang berisi parameter fisiologis dan status merokok. Dataset ini dipilih karena memiliki

jumlah data yang besar, struktur yang jelas, serta relevan dengan tujuan penelitian, yaitu membangun model klasifikasi status perokok berdasarkan indikator fisiologis. Proses akuisisi mencakup pengunduhan dataset, pengecekan format file, serta memastikan bahwa seluruh atribut yang dibutuhkan tersedia dan dapat digunakan untuk proses analisis selanjutnya.

Preprocessing

Tahap preprocessing bertujuan untuk menyiapkan data agar siap digunakan dalam pemodelan. Preprocessing diawali dengan proses Exploratory Data Analysis (EDA) untuk memahami karakteristik data, seperti distribusi variabel, hubungan antar fitur, dan pola umum yang muncul dalam dataset. Setelah itu dilakukan data cleaning yang meliputi pengecekan nilai hilang, pendeteksian dan penghapusan data duplikat, serta identifikasi outlier yang berpotensi memengaruhi stabilitas model. Meskipun outlier dideteksi, nilai-nilai tersebut tidak seluruhnya dihapus karena dianggap sebagai representasi variasi fisiologis yang nyata. Selanjutnya dilakukan tahap data preparation dengan mengubah variabel kategorikal menjadi bentuk numerik melalui proses encoding, serta melakukan standarisasi terhadap fitur numerik untuk menjaga keseragaman skala data.

Splitting Data

Pada tahap ini, dataset dibagi menjadi data latih dan data uji dengan tujuan memisahkan data yang digunakan untuk melatih model dan data yang digunakan untuk menguji kinerja model. Untuk mengatasi permasalahan ketidak-seimbangan jumlah antara kelas perokok dan non-perokok, diterapkan teknik SMOTE (Synthetic Minority Over-sampling Technique) pada data latih. SMOTE bekerja dengan cara menghasilkan sampel sintetis pada kelas minoritas berdasarkan kedekatan data di ruang fitur, sehingga distribusi kelas menjadi lebih seimbang dan model dapat belajar secara lebih adil terhadap kedua kelas.

Training Model

Pada tahap ini, dilakukan pelatihan model menggunakan algoritma Random Forest. Model tidak hanya dilatih dengan parameter default, tetapi juga dioptimasi menggunakan metode Randomized-SearchCV untuk mencari kombinasi parameter terbaik seperti jumlah pohon, kedalaman maksimum, dan parameter lainnya yang memengaruhi performa model. Proses ini bertujuan untuk meningkatkan akurasi dan stabilitas model sekaligus menghindari overfitting terhadap data latih. Model terbaik yang dihasilkan kemudian dipilih untuk digunakan pada tahap evaluasi.

Evaluation

Pada tahap ini, model terbaik diuji menggunakan data uji yang tidak pernah dilihat sebelumnya oleh model selama proses pelatihan. Evaluasi dilakukan menggunakan confusion matrix untuk melihat seberapa baik model dalam mengklasifikasikan kelas perokok dan non-perokok. Selain itu, hasil confusion matrix digunakan sebagai dasar perhitungan metrik evaluasi seperti akurasi, precision, recall, dan F1-score. Tahap evaluasi ini memberikan gambaran objektif mengenai kemampuan model dalam memprediksi status merokok berdasarkan data fisiologis serta menunjukkan tingkat keberhasilan penelitian secara keseluruhan.

RESULTS AND DISCUSSION

Data Akuisisi

Data yang digunakan berasal dari dataset "Smoking Signal of Body Classification" yang diunduh dari platform Kaggle. Dataset ini berisi lebih dari 55.000 data individu dengan 27 atribut fisiologis, seperti tinggi badan, berat badan, tekanan darah sistolik dan diastolik, kadar hemoglobin, kadar kolesterol, GGT, glukosa darah, serta variabel target berupa status merokok.

Tabel 1 Lima baris pertama dataset

ID	gender	age	...	tartar	smoking
0	F	40	...	Y	0
1	F	40	...	Y	0
2	M	55	...	N	1
3	M	40	...	Y	0
4	F	40	...	N	0

Dengan struktur dan jumlah data yang besar serta kelengkapan informasi dalam atribut, dataset ini sangat representatif untuk digunakan dalam proses pelatihan model klasifikasi berbasis machine learning, khususnya untuk mendeteksi perilaku merokok berdasarkan parameter fisiologis individu.

Preprocessing

Tahap preprocessing bertujuan untuk mendapatkan kualitas data yang bagus, karena kualitas data sangat memengaruhi performa model klasifikasi yang dibangun. Preprocessing dilakukan melalui serangkaian Langkah sistematis untuk membersihkan, menyiapkan, dan menyesuaikan data agar sesuai dengan kebutuhan algoritma Random Forest serta meminimalkan potensi kesalahan selama proses pelatihan model.

a. Exploratory Data Analysis (EDA)

EDA dilakukan untuk untuk memahami struktur, tipe data, jumlah entri, dan distribusi variabel target. Dataset yang digunakan dalam penelitian terdiri atas 55.692 baris data dan 26 kolom, yang mencakup atribut numerik (float64 dan int64) serta kategori (object).

```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             55692 non-null  object
 1   age                55692 non-null  int64
 2   height_cm          55692 non-null  int64
 3   weight_kg          55692 non-null  int64
 4   waistcm            55692 non-null  float64
 5   eyesightleft       55692 non-null  float64
 6   eyesightright      55692 non-null  float64
 7   hearingleft        55692 non-null  float64
 8   hearingright       55692 non-null  float64
 9   systolic           55692 non-null  float64
 10  relaxation         55692 non-null  float64
 11  fastingbloodsugar  55692 non-null  float64
 12  cholesterol        55692 non-null  float64
 13  triglyceride       55692 non-null  float64
 14  hdl                55692 non-null  float64
 15  ldl                55692 non-null  float64
 16  hemoglobin         55692 non-null  float64
 17  urineprotein       55692 non-null  float64
 18  serumcreatinine    55692 non-null  float64
 19  ast                55692 non-null  float64
 20  alt                55692 non-null  float64
 21  gtp                55692 non-null  float64
 22  oral               55692 non-null  object
 23  dentalcaries       55692 non-null  int64
 24  tartar             55692 non-null  object
 25  smoking            55692 non-null  int64
dtypes: float64(18), int64(5), object(3)
memory usage: 11.0+ MB
```

Gambar 2 Informasi Dataset

b. Missing Value

Langkah selanjutnya dalam proses pra-pemrosesan data adalah melakukan pengecekan terhadap keberadaan missing value pada setiap kolom fitur dalam dataset. Hal ini penting untuk memastikan bahwa seluruh data yang akan digunakan dalam pelatihan model bersifat lengkap dan tidak mengandung nilai kosong yang dapat mengganggu proses analisis maupun prediksi.

```
--- 4.1.2.2 Pengecekan Missing Value ---
Tidak ditemukan missing value dalam dataset.
```

Gambar 3 Missing Value

c. Data Duplikat

Data duplikat dapat mempengaruhi kualitas pelatihan model karena menyebabkan informasi yang sama dihitung berulang kali, sehingga berpotensi menimbulkan bias pada hasil prediksi.

```
--- 4.1.2.3 Pengecekan Data Duplikat ---
Ditemukan 11140 baris duplikat. Menghapus data duplikat...
Dataset baru setelah menghapus duplikat: (44552, 26)
```

Gambar 4 Data Duplikat

d. Deteksi Outlier

Meskipun keberadaan outlier terdeteksi, dalam penelitian ini tidak dilakukan penghapusan atau penyesuaian terhadap nilai-nilai outlier tersebut. Hal ini karena outlier dianggap masih membawa informasi yang penting secara klinis dan mungkin mencerminkan variasi fisiologis nyata dari responden. Oleh karena itu, seluruh data tetap digunakan dalam tahap pelatihan model agar mencerminkan kondisi dunia nyata secara lebih representatif.

e. Encoding

Proses encoding dilakukan untuk mengubah variabel kategorikal menjadi bentuk numerik agar dapat digunakan dalam model machine learning. Dalam dataset ini, terdapat tiga kolom kategorikal yang di-encode, yaitu: 'gender', 'oral', dan 'tartar'.

Metode encoding yang digunakan adalah label encoding, di mana setiap kategori dalam variabel kategorikal dikonversi menjadi nilai bilangan bulat. Proses ini penting agar algoritma pembelajaran mesin dapat memproses fitur-fitur tersebut secara matematis.

Tabel 2 Encoding

oral	tartar	gender_male
1	1	False
1	1	False
1	0	True
1	1	True
1	0	False

f. Standarisasi

Standarisasi fitur numerik dilakukan untuk menyamakan skala antar fitur, sehingga setiap fitur memiliki kontribusi yang seimbang dalam proses pelatihan model. Teknik standarisasi yang digunakan adalah Z-Score Normalization, yaitu dengan mengurangi nilai fitur dengan rata-rata dan membaginya dengan standar deviasi fitur tersebut.

Tabel 3 Standarisasi

age	height(cm)	...	gender_male
-0.348306	-1.049840	...	-1.322730
-0.348306	-0.506278	...	-1.322730
0.892485	0.580848	...	0.756012
-0.348306	0.037285	...	0.756012
-0.348306	-1.049840	...	-1.322730

Splitting Data

Tahap splitting data bertujuan untuk memisahkan dataset menjadi data pelatihan (training set) dan data pengujian (test set) sehingga model dapat dilatih dan dievaluasi secara independen. Pada penelitian ini, dataset dibagi dengan rasio 80:20, menghasilkan dimensi: X_train: (35,641, 25), X_test: (8,911, 25).

```
--- 4.1.3 Splitting Data ---
Dimensi X_train: (35641, 25)
Dimensi X_test: (8911, 25)
Proporsi kelas di y_train:
```

smoking	proportion
0	63.3%
1	36.7%

Gambar 5 Splitting Data

Distribusi kelas target (smoking) pada data pelatihan awal menunjukkan ketidak-seimbangan data, di mana kelas Non-Smoker mendominasi sebanyak 63.3% dan kelas Smoker hanya 36.7%. Ketidakseimbangan ini berpotensi menurunkan performa model klasifikasi terutama pada prediksi kelas minoritas.

Untuk mengatasi hal tersebut, dilakukan penyeimbangan data menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique). Setelah SMOTE diterapkan, jumlah data pelatihan bertambah menjadi 45.120 sampel, dan proporsi antar kelas menjadi seimbang: Smoker: 50.0%, Non-Smoker: 50.0%.

```
--- Penerapan SMOTE pada Data Training ---
Dimensi X_train setelah SMOTE: (45120, 25)
Proporsi kelas di y_train setelah SMOTE:
```

smoking	proportion
1	50.0%
0	50.0%

Gambar 6 Penerapan SMOTE

Training Model

Pada tahap ini, dilakukan proses optimasi model Random Forest dengan menggunakan pendekatan RandomizedSearchCV. Metode ini bertujuan untuk menemukan kombinasi hyperparameter terbaik yang menghasilkan performa model paling optimal.

```
Randomized Search Selesai.
Hyperparameter Terbaik: {'n_estimators': 200, 'min_samples_split':
Akurasi Model Terbaik Hasil Tuning (pada data test): 0.7638
F1-Score (Perokok/Kelas 1) Terbaik Hasil Tuning: 0.7066
```

Gambar 7 Hyperparameter Tuning

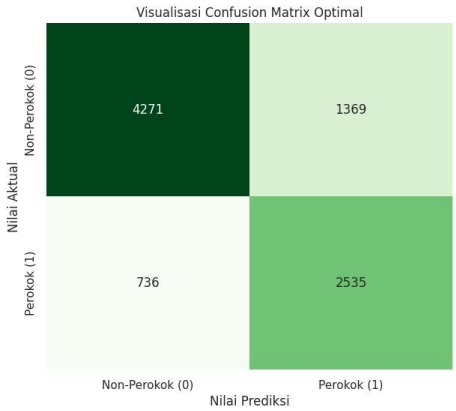
Proses tuning dilakukan dengan pengaturan 3-fold cross-validation untuk setiap 10 kandidat parameter, sehingga menghasilkan total 30 fit. Setelah proses Randomized Search selesai, diperoleh kombinasi hyperparameter terbaik sebagai berikut:

1. n_estimators: 200
2. min_samples_split: 5
3. min_samples_leaf: 1
4. max_depth: None
5. bootstrap: False

Model hasil tuning ini kemudian diuji menggunakan data test dan menghasilkan akurasi sebesar 76.4%.

Evaluation

Evaluasi model dilakukan dengan menggunakan confusion matrix dan classification report. Pada confusion matrix, diketahui bahwa dari 5640 sampel non-perokok, sebanyak 4271 diklasifikasikan dengan benar, sementara 1369 salah diklasifikasikan sebagai perokok. Sebaliknya, dari 3271 sampel perokok, sebanyak 2535 diklasifikasikan dengan benar, dan 736 salah klasifikasi sebagai non-perokok.



Gambar 8 Confusion Matrix

Tabel 4 Classification Report

	precision	recall	f1-score	support
0	0.85	0.76	0.80	5640
1	0.65	0.77	0.71	3271
accuracy			0.76	8911
macro avg	0.75	0.77	0.75	8911
weighted avg	0.78	0.76	0.77	8911

Berdasarkan classification report pada tabel 4 menunjukkan bahwa model memiliki performa yang seimbang antara kedua kelas dengan recall yang cukup tinggi untuk kelas minoritas (perokok). Nilai F1-score sebesar 0.71 untuk kelas perokok mengindikasikan bahwa model cukup baik dalam

mengidentifikasi individu yang merokok, yang merupakan target utama klasifikasi.

CONCLUSION

a. Penerapan algoritma Random Forest dalam penelitian ini dilaksanakan melalui lima tahapan utama yang sistematis. Tahapan ini dirancang untuk memastikan model memiliki performa klasifikasi yang optimal dalam mendeteksi status merokok berdasarkan parameter fisiologis.
b. Kinerja algoritma Random Forest menunjukkan performa yang cukup baik dalam mengklasifikasikan status merokok berdasarkan parameter fisiologis. Setelah dilakukan hyperparameter tuning, model mencapai akurasi sebesar 76.4% pada data uji.
c. Berdasarkan analisis feature importance menggunakan metrik Gini Impurity, diketahui bahwa fitur yang paling berpengaruh dalam menentukan status merokok adalah gender_male, height(cm), hemoglobin, Gtp, dan triglyceride.

REFERENCE

Aishwarya, S., Siddalingaswamy, P. C., & Chadaga, K. (2025). Explainable artificial intelligence driven insights into smoking prediction using machine learning and clinical parameters. *Scientific Reports*, *15*, 24069. https://doi.org/10.1038/s41598-025-09409-w

Ebrahimi, A., Henriksen, M. B. H., Brasen, C. L., Hilberg, O., Hansen, T. F., Jensen, L. H., Peimankar, A., & Wiil, U. K. (2024). Identification of patients' smoking status using an explainable AI approach: a Danish electronic health records case study. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02231-4

Sinha, K., Ghosh, N., & Sil, P. C. (2025). Harnessing machine learning in contemporary tobacco research. In *Toxicology Reports* (Vol. 14). Elsevier Inc. https://doi.org/10.1016/j.toxrep.2024.101877