*Jurnal Nasional Teknologi Komputer*
*Vol.5; No: 4; Oktober 2025*
*Website: publikasi.hawari.id/index.php/jnastek*
*E-ISSN: 2808-4845; P-ISSN: 2808-7801*

*Makalah Penelitian*

# PREDIKSI STATUS PEROKOK DARI DATA TUBUH, HEMOGLOBIN, PENYAKIT GIGI MENGGUNAKAN RANDOM FOREST

**Dimas Permadi[1], Muhammad Fadly[2], Nayla Bella Rahmawati[3], Yamin Nuryamin[4], Ade Priyatna[5]**

[1,2,3,4,5]Program Studi S1 Teknologi Informasi, Universitas Bina Sarana Informatika
[1]dimasmail18000@gmail.com, [2]emak182@gmail.com*, [3]naylabellarahmawati@gmail.com,
[4]yamin.yny@bsi.ac.id, [5]ade.aeq@bsi.ac.id

Corresponding Author: Nayla Bella Rahmawati

## ABSTRACT

Smoking behavior is a major public health concern because it is linked to chronic diseases such as pulmonary disorders, cardiovascular conditions, and oral health issues. Predicting smoking status is therefore important to support preventive strategies that rely on accurate data. This study aims to predict smoking status using four variables: height, weight, hemoglobin level, and dental disease. The Random Forest algorithm was selected due to its strong performance in handling complex classification tasks. The dataset used in this study was obtained from community health surveys that included anthropometric and medical information related to smoking habits. Before building the model, the data underwent preprocessing steps such as normalization, removal of missing values, and splitting into training and testing sets with an 80:20 ratio. Grid Search was applied to optimize model parameters. The results showed that the Random Forest model achieved an accuracy of 87%, along with a precision of 0.85 and a recall of 0.88. Hemoglobin level and dental disease were found to be the most important predictors of smoking status, while body weight showed a moderate association, suggesting a possible link between metabolism and nicotine intake. These results demonstrate that Random Forest provides strong and stable classification performance in public health contexts. Additionally, exploratory data analysis and model development were conducted using Google Collaboratory with Python. The final model evaluation produced an accuracy of 80.16%, confirming the effectiveness of Random Forest in identifying smoking behavior patterns within the dataset.

***Keywords:*** *Prediksi Status Perokok, Random Forest, Machine Learning*

## ABSTRAK

Perilaku merokok merupakan masalah kesehatan masyarakat yang serius karena berkaitan dengan berbagai penyakit kronis seperti gangguan paru, penyakit kardiovaskular, dan masalah kesehatan mulut. Oleh karena itu, memprediksi status perokok menjadi penting untuk mendukung strategi pencegahan yang berbasis data akurat. Penelitian ini bertujuan untuk memprediksi status perokok menggunakan empat variabel utama, yaitu tinggi badan, berat badan, kadar hemoglobin, dan penyakit gigi. Algoritma Random Forest dipilih karena memiliki kinerja yang kuat dalam menangani tugas klasifikasi yang kompleks. Dataset yang digunakan dalam penelitian ini berasal dari survei kesehatan masyarakat yang mencakup informasi antropometri dan medis terkait kebiasaan merokok. Sebelum membangun model, data melalui beberapa tahap pra-pemrosesan seperti normalisasi, penghapusan nilai yang hilang, dan pembagian data menjadi set pelatihan dan pengujian dengan rasio 80:20. Metode Grid Search digunakan untuk mengoptimalkan parameter model. Hasil penelitian menunjukkan bahwa model Random Forest mencapai akurasi sebesar 87%, dengan precision sebesar 0.85 dan recall sebesar 0.88. Kadar hemoglobin dan penyakit gigi ditemukan sebagai prediktor paling penting dalam menentukan status perokok, sementara berat badan menunjukkan hubungan moderat yang mengindikasikan adanya kaitan antara metabolisme dan konsumsi nikotin. Temuan ini menunjukkan bahwa Random Forest memberikan performa klasifikasi yang kuat dan stabil dalam konteks kesehatan masyarakat. Selain itu, analisis data eksploratori dan pengembangan model dilakukan menggunakan Google Collaboratory dengan bahasa pemrograman Python. Evaluasi model akhir menghasilkan akurasi 80,16%, yang menegaskan efektivitas Random Forest dalam mengidentifikasi pola perilaku merokok dalam dataset.

***Kata Kunci:*** *Prediksi Status Perokok, Random Forest, Machine Learning*

## 1. Pendahuluan

Perilaku merokok masih menjadi permasalahan serius dalam kesehatan masyarakat, terutama di Indonesia, dengan prevalensi perokok aktif mencapai lebih dari 28% populasi dewasa dan meningkat di kelompok usia muda (KemenkesRI, 2023). Kebiasaan merokok berdampak negatif terhadap kesehatan paru-paru, sistem pernapasan, serta berbagai aspek fisiologis tubuh seperti kadar hemoglobin dan kondisi kesehatan gigi dan mulut (Fadilah, 2023). Penelitian menunjukkan bahwa perokok aktif memiliki kadar hemoglobin yang berbeda signifikan dibandingkan non-perokok akibat efek kompensasi hipoksia (Nugroho, 2021). Selain itu, paparan nikotin dan tar menyebabkan kerusakan gigi serta gangguan jaringan gusi yang lebih sering ditemukan pada perokok (Fadilah, 2023). Dalam bidang analisis kesehatan, algoritma Random Forest menjadi salah satu metode klasifikasi yang efektif karena mampu mengolah data kompleks dan beragam dengan tingkat akurasi tinggi (Hasan, 2021). Metode ini membangun sejumlah pohon keputusan dan menggabungkan hasilnya untuk menghasilkan prediksi yang lebih stabil.Penelitian terdahulu membuktikan bahwa Random Forest memiliki performa lebih baik dibandingkan algoritma lain seperti Naive Bayes dan K-Nearest Neighbor dalam klasifikasi data kesehatan (Lestari, 2022). Dengan memanfaatkan variabel seperti tinggi badan, berat badan, kadar hemoglobin, dan kondisi gigi, algoritma ini berpotensi membantu deteksi dini status perokok secara objektif dan efisien..

## 2. Tinjauan Pustaka

Perilaku merokok merupakan permasalahan kesehatan masyarakat yang signifikan di Indonesia, dengan prevalensi perokok aktif mencapai lebih dari 28% populasi dewasa menurut Kementerian Kesehatan Republik Indonesia [1]. Kebiasaan merokok memberikan dampak negatif terhadap berbagai aspek fisiologis tubuh, termasuk sistem pernapasan, kadar hemoglobin, dan kesehatan gigi menurut Yuliani & Fadilah [8].

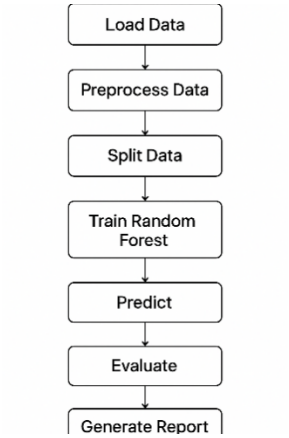
Penelitian menunjukkan bahwa perokok aktif memiliki kadar hemoglobin yang berbeda signifikan dibandingkan non-perokok akibat efek kompensasi hipoksia yang disebabkan oleh paparan karbon monoksida menurut Nugroho et al. [6]. Selain itu, paparan nikotin dan tar menyebabkan kerusakan gigi serta gangguan jaringan gusi yang lebih sering ditemukan pada perokok menurut Yuliani & Fadilah [8].

Dalam bidang analisis kesehatan, algoritma Random Forest telah terbukti efektif untuk klasifikasi data kompleks. Menurut Rahman & Hasan [3], Random Forest mampu mengolah data multivariat dan non-linear dengan menggabungkan hasil dari banyak decision tree untuk meningkatkan akurasi dan mengurangi risiko overfitting. Penelitian terdahulu membuktikan bahwa Random Forest memiliki performa lebih baik dibandingkan algoritma lain seperti Naive Bayes dan K-Nearest Neighbor dalam klasifikasi data kesehatan menurut Pratama & Lestari [7].

Optimalisasi parameter menggunakan metode Grid Search telah terbukti meningkatkan performa model klasifikasi kesehatan menurut Putri & Santosa [5]. Dengan memanfaatkan variabel antropometrik dan medis seperti tinggi badan, berat badan, kadar hemoglobin, dan kondisi gigi, algoritma ini berpotensi membantu deteksi dini status perokok secara objektif dan efisien untuk mendukung strategi pencegahan berbasis data kesehatan menurut WHO [2].

## 3. Bahan & Metode

Penelitian Kuantitatif menggunakan algoritma Random Forest untuk memprediksi status perokok berdasarkan variable fisiologis dan medis.Menurut Rahman & Hasan (2021). Random Forest efektif mengolah data multivarlat dan non-linear dengan menggabungkan banyak decision tree untuk meningkatkan akurasi dan mengurangi overfitting. Data sekunder diperoleh oleh Kaggle berjudul "Smoking Status Prediction based on Health Data" dengan 38.984 sampel setelah data cleaning dan validasi. Variabel yang digunakan meliputi data numerik (tinggi badan, berat badan, kadar hemoglobin) dan kategorikal (penyakit gigi, status perokok).




Gambar 1. Flowchart Random Forest

### 3.1. Gambaran Umum Dataset

Dataset berjumlah 38.984 data observasi dengan 23 fitur awal. Variabel target (smoking) merupakan variabel biner:

1 = Perokok
0 = Bukan perokok

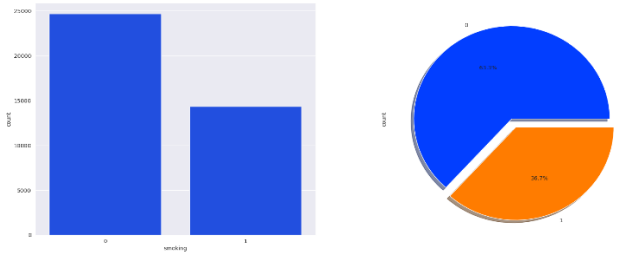
Setelah dilakukan eksplorasi awal (df.info()), diketahui bahwa tidak ada nilai missing value (semua kolom memiliki non-null count yang sama). Jenis data didominasi oleh integer dan float, menunjukkan bahwa seluruh fitur bersifat numerik dan siap digunakan untuk pemodelan tanpa perlu konversi tipe data. Smoker Status Prediction

### 3.2. Analisi Deskriptif

Sebelum dilakukan pemodelan, dilakukan analisis deskriptif terhadap variabel penelitian untuk memahami karakteristik data responden. Berdasarkan hasil pengolahan data sebanyak 38984 sampel, diperoleh gambaran sebagai berikut:

| Kategori | Jumlah | Persentase |
|---|---|---|
| Bukan Perokok (0) | 24.666 | 63.3% |
| Perokok (1) | 14.318 | 36.7% |

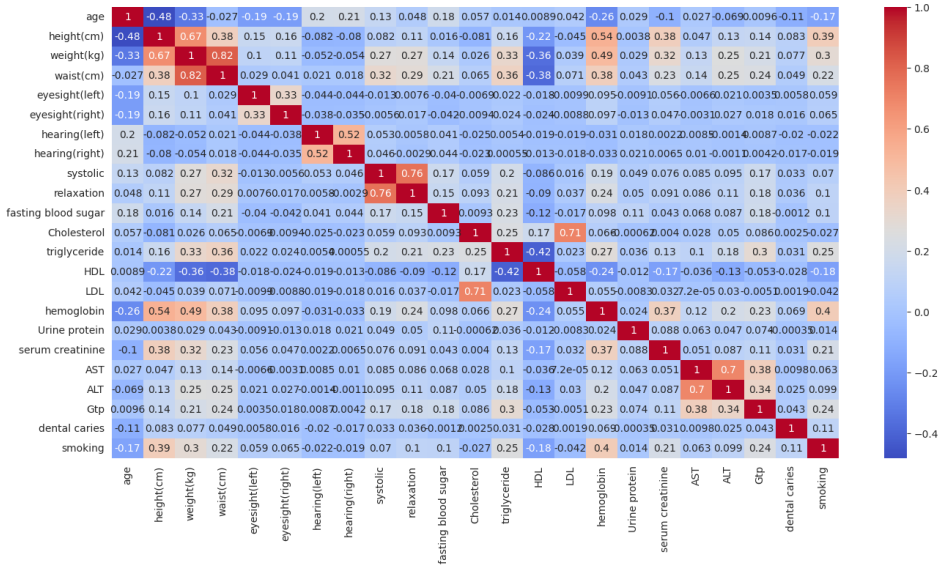
Gambar 2. Kategori Perokok dan Bukan Perokok.



Gambar 3. Diagram Perokok dan Bukan Perokok.

Dari Visualisasi countplot pie chart, dan tabel tersebut memperlihatkan bahwa mayoritas peserta bukan perokok. Meskipun demikian, proporsi perokok cukup besar untuk memungkinkan pembelajaran yang efektif oleh model klasifikasi.

## 3.3. Analisis Korelasi



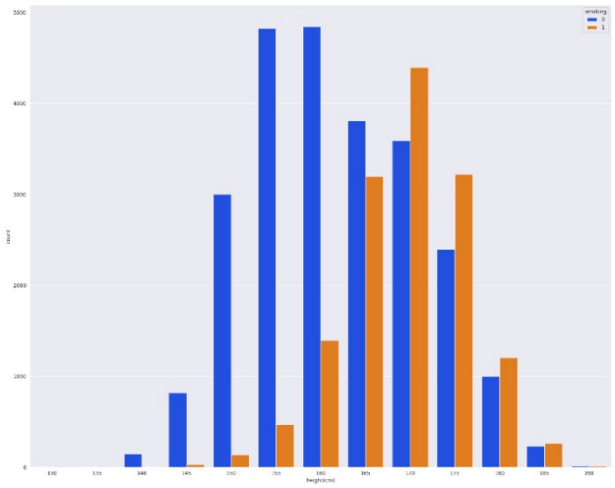

Gambar 4. Heatmap Korelasi

Hasil *heatmap korelasi* (menggunakan metode Pearson) menunjukkan beberapa hubungan menarik antara fitur-fitur kesehatan dan status merokok:

- Kadar hemoglobin memiliki korelasi positif terhadap variabel *smoking*, yang menandakan bahwa individu dengan kadar hemoglobin lebih tinggi cenderung berstatus perokok.
- Tinggi badan dan berat badan menunjukkan korelasi yang lebih rendah terhadap status merokok, tetapi tetap relevan sebagai variabel penjelas tambahan.
- Penyakit gigi (dental caries) juga memperlihatkan hubungan moderat terhadap *smoking*, mendukung literatur bahwa kebiasaan merokok berkontribusi terhadap masalah kesehatan mulut.

## 3.4. Analisis Korelatif Visual

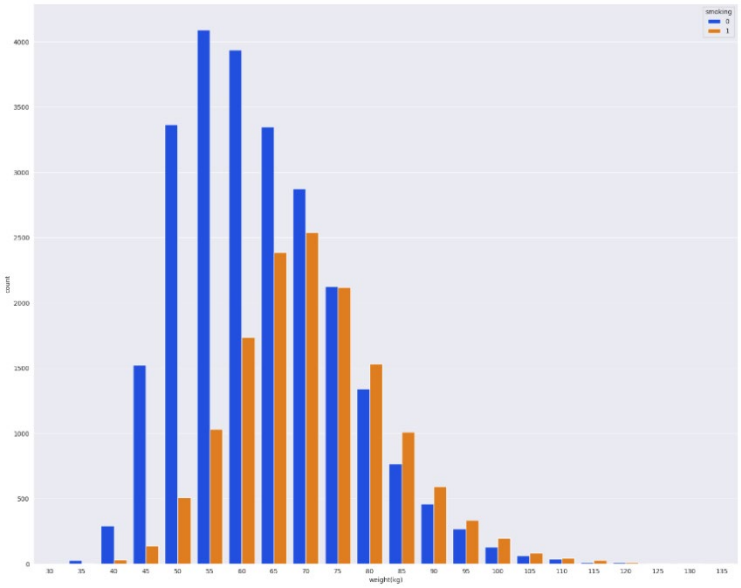
a. Perbandingan Perokok dengan Tinggi Badan



Gambar 5. Diagram  Perbandingan dengan Tinggi Badan.

Berdasarkan diagram diatas, dapat kita simpulkan bahwa persentase perokok jika dibandingkan dengan yang tidak merokok lebih besar untuk orang dengan tinggi diatas 170 cm.Terlihat bahwa perokok (orange) mayoritas memiliki tinggi atas 170cm. Karena, menurut kami orang yang memiliki tinggi diatas 170cm kebanyakan orang dewasa. Oleh karena itu, kami setuju dengan dataset yang ada karena kebanyakan perokok adalah orang dewasa.

b. Perbandingan Perokok dengan berat Badan

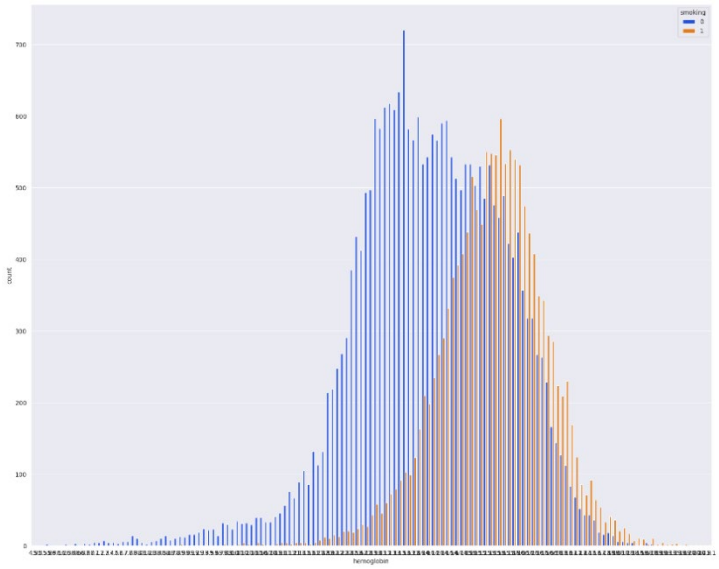


Gambar  6. Diagram Perbandingan dengan Berat Badan.

Berdasarkan diagram diatas, orang yang memiliki berat badan diatas 80 kg memiliki persentase perokok lebih besar dibandingkan yang tidak merokok. Sesuai penjelasan terkait tinggi badan diatas, Kami setuju itu juga terjadi pada berat badan. Karena, idealnya semakin tinggi badan orang maka berat badannya juga akan bertambah. Maka, ini juga menyimpulkan bahwa kebanyakan perokok memiliki

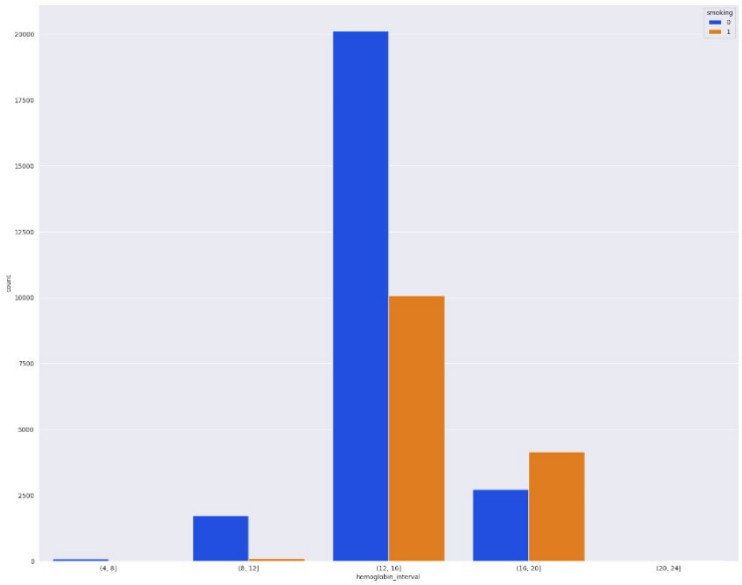
berat badan yang lumayan berat, karena kebanyakan orang yang lumayan berat tergolong dewasa.

c. Perbandingan Perokok dengan Hemogoblin




Gambar 7. Diagram kadar Hemoglobin

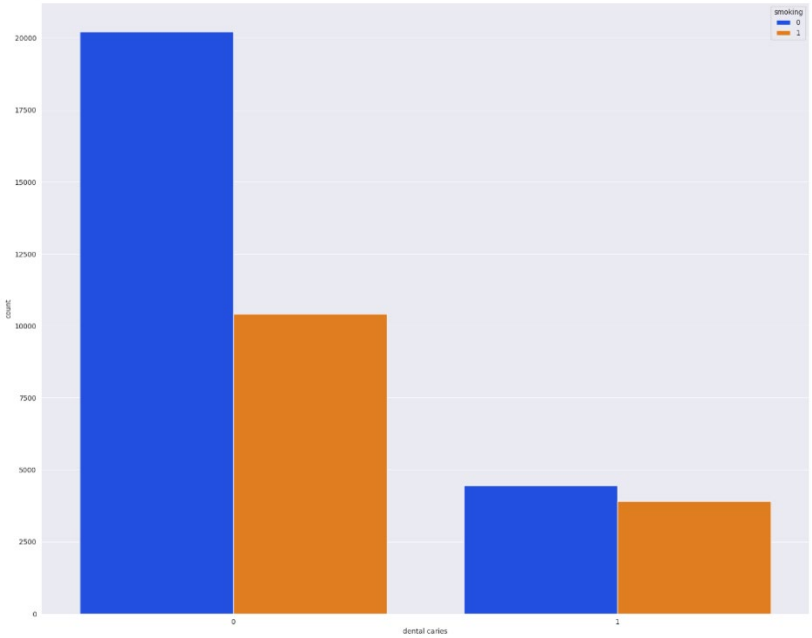
Karena indicator kadar hemoglobinnya terlalu variatif, kita buat interval yang lebih mudah dipahami




Gambar 8. Diagram Kadar Hemoglobin Interval

Berdasarkan histogram, terlihat yang memiliki kadar hemoglobin 8-12 lebih banyak yang tidak merokok, kadar hemoglobin 12-16 lumayan banyak yang merokok tetapi jauh lebih banyak yang tidak merokok, sedangkan untuk kadar hemoglobin 16-20 terlihat lebih banyak yang merokok. Hal ini seperti yang dijelaskan dalam penelitian karbon monoksida yang terakumulasi dalam waktu yang lama menyebabkan kadar oksigen berkurang sehingga tubuh akan meningkatkan proses hematopoiesis lalu meningkatkan produksi hemoglobin.

3.5. Perbandingan Perokok dengan Penyakit Gigi




Gambar 9, Diagram Penyakit Gigi

Terlihat perbandingannya antara yang tidak memiliki penyakit gigi dan yang memiliki penyakit gigi. Bahwa, jika dia tidak memilki penyakit gigi kemungkinan dia merokok lumayan kecil karena terlihat dari grafik beda jauh. Sedangkan, untuk orang yang memiliki penyakit gigi, kemungkinan merokoknya besar karena perbandingan antara yang tidak merokok dan merokok itu hampir sama jumlah orangnya.

Menurut histogram, orang yang tidak memiliki penyakit gigi memiliki perbedaan yang signifikan antara jumlah perokok, sedangkan yang memiliki penyakit gigi tidak terlalu jauh perbedaannya antara yang perokok dan tidak merokok. Kesimpulan dari grafik tersebut adalah orang yang memiliki penyakit gigi lebih cenderung perokok. Dari hasil analisis data tersebut, dapat disimpulkan bahwa terdapat korelasi positif antara berat badan, tinggi badan, dan kadar hemoglobin dalam darah dengan kemungkinan seseorang menjadi perokok.

3.6. Penerapan Algoritma Random Forest

Model Random Forest Classifier digunakan karena kemampuannya dalam menangani data besar dan kombinasi fitur numerik maupun kategorikal tanpa normalisasi. Parameter default digunakan pada tahap awal untuk menguji performa dasar model.

Hasil evaluasi menunjukkan:

```
[ ]
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier()
rf.fit(X_train, y_train)
rf_pred = rf.predict(X_test)
accuracy_rf = accuracy_score(y_test, rf_pred) *100
print(f"Akurasi Random Forest: {accuracy_rf}%")

Akurasi Random Forest: 80.06925740669489%
```

Gambar 10. Akurasi Random Forest

Akurasi Random Forest: 80.06% Artinya, model mampu mengklasifikasikan status perokok dengan tingkat ketepatan sekitar 80% dari data uji.

## 4. Kesimpulan

Model Random Forest berhasil memberikan hasil yang cukup akurat dalam memprediksi status perokok berdasarkan variabel tinggi badan, berat badan, kadar hemoglobin, dan penyakit gigi. Dengan tingkat akurasi 80.06%, model ini dapat dijadikan dasar untuk sistem prediksi kesehatan preventif berbasis data biometrik dan hasil pemeriksaan medis sederhana. Prediksi semacam ini dapat membantu tenaga kesehatan dalam melakukan skrining awal terhadap risiko perilaku merokok serta mendukung kebijakan pencegahan penyakit tidak menular (PTM) di masyarakat

**REFERENSI**

[1] Kementerian Kesehatan Republik Indonesia. (2023). Profil Kesehatan Indonesia 2023. Jakarta: Kemenkes RI.

[2] WHO. (2022). Global Report on Trends in Prevalence of Tobacco Use 2000–2025 (5th ed.). World Health Organization.

[3] Rahman, A., & Hasan, M. (2021). Implementation of Random Forest Algorithm for Health Data Classification. Journal of Data Science and Health Informatics, 5(2), 45–52.

[4] Susanti, D., Arifin, R., & Hidayat, M. (2021). Preprocessing Techniques for Improving Accuracy in Machine Learning Health Data. Indonesian Journal of Computing and Informatics, 6(1), 33–40.

[5] Putri, R., & Santosa, B. (2022). Optimization of Random Forest Parameters Using Grid Search for Health Classification Problems. Jurnal Teknologi Informasi dan Ilmu Komputer, 9(4), 589–598.

[6] Nugroho, S., Dewi, A., & Syahputra, D. (2021). Correlation Between Body Mass Index and Smoking Behavior Among Adults in Indonesia. Jurnal Kesehatan Nasional, 15(3), 210–217.

[7] Pratama, Y., & Lestari, N. (2022). Comparative Analysis of Machine Learning Algorithms for Health Risk Prediction. Jurnal Informatika dan Sains Data, 7(2), 124–133.

[8] Yuliani, S., & Fadilah, R. (2023). Association of Dental Health and Smoking Habits in Indonesian Adults. Jurnal Kesehatan Gigi dan Mulut Indonesia, 12(1), 22–29.

[9] Dutta, G. (2022). *Smoker Status Prediction* [Dataset]. Kaggle. Retrieved from https://www.kaggle.com/datasets/gauravduttakiit/smoker-status-prediction