**BULLET : Jurnal Multidisiplin Ilmu**
Volume 2, No. 02, April - Mei 2023
ISSN 2829-2049 (media online)
Hal 568-573

# Perbandingan Akurasi Algoritma *Random Forest* Dan *Naïve Bayes* Dalam Memprediksi Risiko Hipertensi

**Tati Suprapti[1], Saeful Anwar[2]***

[1,2]Program Studi Teknik Informatika, STMIK IKMI Cirebon, Kota Cirebon, Indonesia
Email: [1]tatisuprapti.ikmi@gmail.com, [2*]saefulanwar.ikmi@gmail.com
(* : coressponding author)

**Abstrak** − Penyakit hipertensi adalah salah satu penyebab utama dari berbagai komplikasi kesehatan yang serius. Oleh karena itu, sangat penting untuk dapat memprediksi risiko hipertensi sedini mungkin agar dapat diambil langkah preventif. Penelitian ini bertujuan untuk membandingkan akurasi dua algoritma machine learning, yaitu *Random Forest* dan *Naïve Bayes*, dalam memprediksi risiko hipertensi menggunakan dataset yang berisi informasi tentang faktor-faktor yang mempengaruhi kesehatan. Kedua algoritma tersebut diterapkan untuk mengklasifikasikan data pasien ke dalam dua kategori: risiko hipertensi tinggi dan risiko hipertensi rendah. Berdasarkan pengujian menggunakan beberapa metrik evaluasi seperti akurasi, precision, recall, dan F1-score, hasilnya menunjukkan bahwa algoritma *Random Forest* memberikan hasil yang lebih baik dibandingkan *Naïve Bayes*, dengan tingkat akurasi yang lebih tinggi dan performa yang lebih konsisten. Temuan ini dapat digunakan sebagai referensi dalam pengembangan sistem pendukung keputusan untuk deteksi dini hipertensi di masyarakat.

**Kata Kunci:** *Random Forest, Naïve Bayes*, Prediksi Risiko Hipertensi, Algoritma, Akurasi

***Abstract*** *− Hypertension is one of the leading causes of serious health complications. Therefore, it is crucial to predict hypertension risks early to take preventive measures. This study aims to compare the accuracy of two machine learning algorithms, Random Forest and Naïve Bayes, in predicting hypertension risks using a dataset containing information about factors affecting health. Both algorithms were applied to classify patient data into two categories: high hypertension risk and low hypertension risk. Based on testing using evaluation metrics such as accuracy, precision, recall, and F1-score, the results showed that the Random Forest algorithm performed better than Naïve Bayes, with higher accuracy and more consistent performance. This finding can be used as a reference for the development of a decision support system for early hypertension detection in the community.*

***Keywords****: Random Forest, Naïve Bayes, Hypertension Risk Prediction, Algorithm, Accuracy*

## 1. PENDAHULUAN

Hipertensi, atau tekanan darah tinggi, merupakan penyakit yang terjadi ketika tekanan darah di dalam arteri meningkat secara konsisten. Penyakit ini menjadi faktor risiko utama untuk penyakit jantung, stroke, dan gagal ginjal. Oleh karena itu, prediksi dini risiko hipertensi sangat penting untuk mengambil langkah-langkah pencegahan yang tepat. Saat ini, prediksi hipertensi tidak hanya bergantung pada pemeriksaan fisik, tetapi juga dapat dilakukan menggunakan algoritma machine learning, yang dapat memanfaatkan data historis pasien untuk memprediksi risiko penyakit.

Dalam penelitian ini, dua algoritma yang sangat populer dalam klasifikasi data yaitu *Random Forest* dan *Naïve Bayes* dibandingkan untuk mengevaluasi kemampuan mereka dalam memprediksi risiko hipertensi. *Random Forest*, yang merupakan algoritma ensemble learning, sering kali digunakan dalam pengolahan data besar dan kompleks. Sementara itu, *Naïve Bayes*, meskipun lebih sederhana, juga banyak digunakan dalam analisis klasifikasi karena efisiensinya dalam memproses data dengan jumlah fitur yang besar.

Metode *Random Forest* bekerja dengan membangun sejumlah decision trees dan menggabungkan hasilnya untuk meningkatkan akurasi klasifikasi. Sebaliknya, *Naïve Bayes* menggunakan prinsip probabilitas untuk memodelkan hubungan antar fitur dalam dataset. Kedua algoritma ini telah banyak digunakan dalam berbagai bidang kesehatan untuk memprediksi berbagai kondisi medis, termasuk hipertensi.

Penelitian ini bertujuan untuk membandingkan akurasi kedua algoritma tersebut dalam memprediksi risiko hipertensi dan memberikan rekomendasi terkait algoritma mana yang lebih cocok untuk aplikasi dalam sistem prediksi kesehatan berbasis data..

**BULLET : Jurnal Multidisiplin Ilmu**
Volume 2, No. 02, April - Mei 2023
ISSN 2829-2049 (media online)
Hal 568-573

# 2. METODOLOGI PENELITIAN

Metodologi dalam penelitian ini terdiri dari beberapa tahapan utama, yaitu pengumpulan data, preprocessing data, penerapan algoritma *Random Forest* dan *Naïve Bayes*, serta evaluasi model. Berikut adalah penjelasan rinci untuk setiap tahapan yang dilakukan dalam penelitian ini.

## 2.1 Pengumpulan Data

Data yang digunakan berasal dari berbagai sumber dataset kesehatan yang mencakup informasi demografis dan medis pasien. Dataset ini mencakup lebih dari 1.000 pasien dengan berbagai fitur terkait risiko hipertensi. Data ini mencakup informasi berikut:

a) Usia

Usia pasien dalam tahun. Risiko hipertensi meningkat dengan bertambahnya usia.

b) Jenis Kelamin

Jenis kelamin pasien, karena pria dan wanita memiliki pola risiko yang berbeda.

c) Indeks Massa Tubuh (BMI)

Indeks yang digunakan untuk menilai status berat badan, dengan BMI tinggi terkait dengan hipertensi.

d) Riwayat Keluarga

Adanya riwayat hipertensi dalam keluarga yang dapat meningkatkan risiko individu.

e) Tekanan Darah

Pengukuran tekanan darah yang dapat menunjukkan status hipertensi atau risiko hipertensi.

f) Gaya Hidup (Pola makan, aktivitas fisik)

Pola makan dan tingkat aktivitas fisik yang mempengaruhi kesehatan pembuluh darah dan tekanan darah

**Tabel 1** Fitur Data pada Dataset Hipertensi

| Fitur | Deskripsi |
|---|---|
| Usia | Usia pasien dalam tahun |
| Jenis Kelamin | Jenis kelamin pasien |
| Indeks Massa Tubuh | Nilai BMI pasien |
| Riwayat Keluarga | Riwayat hipertensi keluarga |
| Tekanan Darah | Nilai tekanan darah pasien |

## 2.2. Preprocessing Data

Data yang dikumpulkan perlu melalui beberapa tahapan preprocessing agar dapat digunakan dalam algoritma *Random Forest* dan *Naïve Bayes*. Proses ini meliputi langkah-langkah berikut:

a) Pembersihan Data: Menghapus data yang hilang atau tidak lengkap.

**BULLET : Jurnal Multidisiplin Ilmu**
Volume 2, No. 02, April - Mei 2023
ISSN 2829-2049 (media online)
Hal 568-573

b) Transformasi Data: Mengubah data yang bersifat kategorikal menjadi numerik, seperti mengonversi jenis kelamin (laki-laki/perempuan) menjadi angka.

c) Pembagian Data: Dataset dibagi menjadi dua bagian: training set (80%) dan testing set (20%) untuk evaluasi model.

### 2.3 Implementasi Algoritma *Random Forest* dan *Naïve Bayes*

Setelah preprocessing, data digunakan untuk membangun model *Random Forest* dan *Naïve Bayes*. Kedua algoritma ini diterapkan untuk memprediksi risiko hipertensi berdasarkan data yang ada.

a) *Random Forest*: Dikenal karena kemampuannya untuk mengurangi overfitting dengan membangun banyak decision trees. Setiap tree mengklasifikasikan data dan hasilnya digabungkan untuk menghasilkan prediksi akhir.

b) *Naïve Bayes*: Berdasarkan teori probabilitas dan menggunakan asumsi bahwa fitur-fitur dalam data bersifat independen satu sama lain, meskipun dalam kenyataannya mungkin tidak selalu demikian.

### 2.4 Evaluasi Model

Untuk mengevaluasi hasil klasifikasi, kami menggunakan beberapa metrik performa berikut:

a) Akurasi: Mengukur sejauh mana prediksi yang benar dibandingkan dengan total data.

b) Precision: Mengukur ketepatan model dalam mengklasifikasikan data positif.

c) Recall: Mengukur sensitivitas model dalam mendeteksi kasus positif.

d) F1-score: Rata-rata harmonis dari precision dan recall, memberikan gambaran umum tentang keseimbangan model.

## 3. ANALISA DAN PEMBAHASAN

Pada bagian ini, kami akan membahas hasil yang diperoleh dari penerapan algoritma *Random Forest* dan *Naïve Bayes* dalam memprediksi risiko hipertensi. Selain itu, kami juga akan menginterpretasikan hasil evaluasi yang dilakukan dengan berbagai metrik performa, seperti akurasi, precision, recall, dan F1-score, untuk menganalisis kekuatan dan kelemahan masing-masing algoritma.

### 3.1 Hasil Klasifikasi dengan *Random Forest* dan *Naïve Bayes*

Setelah menerapkan *Random Forest* dan *Naïve Bayes* pada data yang telah diproses, kami memperoleh hasil klasifikasi yang menunjukkan seberapa baik kedua algoritma ini dalam memprediksi risiko hipertensi berdasarkan data yang tersedia.

Hasil evaluasi menggunakan metrik akurasi dan metrik lainnya dapat dilihat pada Tabel 2 dan Tabel 3.

**Tabel 2** Metrik Evaluasi untuk *Random Forest*

| Metrik | Nilai |
|---|---|
| Akurasi | 91% |
| Precision | 92% |
| Recall | 89% |
| F1-Score | 90.5% |

**BULLET : Jurnal Multidisiplin Ilmu**
Volume 2, No. 02, April - Mei 2023
ISSN 2829-2049 (media online)
Hal 568-573

**Tabel 3** Metrik Evaluasi untuk *Naïve Bayes*

| Metrik | Nilai |
|---|---|
| Akurasi | 85% |
| Precision | 87% |
| Recall | 84% |
| F1-Score | 85.5% |

Dari tabel di atas, dapat dilihat bahwa *Random Forest* menghasilkan nilai akurasi yang lebih tinggi dibandingkan dengan *Naïve Bayes*. Hal ini menunjukkan bahwa *Random Forest* lebih unggul dalam menangani dataset dengan banyak fitur dan kompleksitas yang tinggi.

### 3.2 Analisis Kinerja *Random Forest*

*Random Forest* menggabungkan hasil dari banyak decision trees untuk meningkatkan akurasi klasifikasi. Keunggulan utama dari *Random Forest* adalah kemampuannya untuk menangani dataset yang besar dan kompleks dengan lebih baik, serta mengurangi kemungkinan overfitting yang sering terjadi pada model yang lebih sederhana. Dalam kasus ini, *Random Forest* berhasil mengidentifikasi pola-pola yang lebih kompleks dalam data pasien, seperti hubungan antara BMI, usia, jenis kelamin, dan riwayat keluarga dalam menentukan risiko hipertensi.

Keakuratan yang tinggi dari *Random Forest* juga didukung oleh kemampuannya dalam mengatasi ketergantungan antar fitur. Misalnya, hubungan antara tekanan darah dan riwayat keluarga mungkin lebih sulit dikenali oleh algoritma yang lebih sederhana, tetapi *Random Forest* mampu menangani fitur-fitur yang saling berinteraksi dan menghasilkan prediksi yang lebih akurat.

### 3.3 Analisis Kinerja *Naïve Bayes*

Sementara itu, *Naïve Bayes* cenderung memiliki performa yang lebih rendah dibandingkan *Random Forest* dalam penelitian ini. Hal ini kemungkinan disebabkan oleh asumsi independensi antara fitur-fitur dalam *Naïve Bayes*, yang berarti algoritma ini menganggap bahwa semua fitur tidak saling bergantung satu sama lain. Pada kenyataannya, dalam dataset hipertensi, banyak fitur yang saling berinteraksi, seperti indeks massa tubuh (BMI) yang berhubungan dengan tekanan darah, gaya hidup, dan riwayat keluarga.

Meski demikian, *Naïve Bayes* tetap memberikan hasil yang baik, terutama pada data yang lebih sederhana. Keuntungan utama dari *Naïve Bayes* adalah kecepatan dan efisiensinya dalam melakukan klasifikasi, sehingga cocok digunakan dalam aplikasi yang membutuhkan waktu komputasi yang singkat.

### 3.4 Pebandingan Antara *Random Forest* dan *Naïve Bayes*

Berdasarkan hasil yang diperoleh, dapat disimpulkan bahwa *Random Forest* lebih unggul dalam menangani dataset yang kompleks dan besar, serta memberikan hasil klasifikasi yang lebih akurat. *Naïve Bayes*, meskipun lebih sederhana dan cepat, kurang efektif dalam menangani interaksi antar fitur yang kompleks. Hal ini membuat *Random Forest* lebih cocok digunakan dalam prediksi risiko hipertensi yang melibatkan banyak variabel.

Namun, *Naïve Bayes* tetap dapat menjadi pilihan yang baik dalam kasus di mana data tidak terlalu kompleks dan waktu komputasi menjadi pertimbangan utama. *Naïve Bayes* juga lebih mudah untuk diimplementasikan dan lebih efisien dalam hal penggunaan memori, yang membuatnya berguna dalam aplikasi dengan sumber daya terbatas.

**BULLET : Jurnal Multidisiplin Ilmu**
Volume 2, No. 02, April - Mei 2023
ISSN 2829-2049 (media online)
Hal 568-573

**Tabel 4** Perbandingan Akurasi Metrik Antara *Random Forest* dan *Naïve Bayes*

| Metrik | *Random Forest* | *Naïve Bayes* |
|---|---|---|
| Akurasi | 91% | 85% |
| Precision | 92% | 87% |
| Recall | 89% | 84% |
| F1-Score | 90.5% | 85.5% |

Keterangan: Tabel ini menunjukkan perbandingan hasil evaluasi performa kedua algoritma dalam memprediksi risiko hipertensi.

### 3.5 Implikasi untuk Pengeolaan Kesehatan

Hasil dari analisis ini memiliki implikasi penting dalam pengelolaan sistem kesehatan berbasis data mining. *Random Forest* dapat digunakan untuk mengembangkan sistem pendukung keputusan yang lebih akurat dalam memprediksi risiko hipertensi pada populasi yang lebih besar. Dengan menggunakan model ini, profesional medis dapat mengidentifikasi pasien yang berisiko tinggi untuk mengalami hipertensi dan memberikan perawatan atau intervensi lebih dini.

Di sisi lain, *Naïve Bayes* dapat digunakan sebagai alternatif yang lebih efisien untuk aplikasi dengan sumber daya terbatas, meskipun dengan tingkat akurasi yang sedikit lebih rendah. Penggunaan *Naïve Bayes* cocok dalam skenario yang membutuhkan model cepat dan ringan, misalnya pada perangkat mobile atau aplikasi yang memerlukan analisis cepat

## 4. KESIMPULAN

Penelitian ini membandingkan algoritma *Random Forest* dan *Naïve Bayes* dalam memprediksi risiko hipertensi menggunakan dataset kesehatan. Hasil penelitian menunjukkan bahwa *Random Forest* menghasilkan akurasi yang lebih tinggi dibandingkan *Naïve Bayes*, serta performa yang lebih konsisten dalam klasifikasi risiko hipertensi. Oleh karena itu, *Random Forest* lebih direkomendasikan untuk digunakan dalam prediksi risiko hipertensi di lingkungan medis.

Penelitian ini dapat dikembangkan lebih lanjut dengan melibatkan metode machine learning lainnya, seperti Deep Learning, untuk membandingkan akurasi dan efektivitas model dalam skala yang lebih besar. Selain itu, penelitian lebih lanjut juga dapat mempertimbangkan faktor-faktor lain yang mempengaruhi prediksi hipertensi, seperti faktor lingkungan dan pola hidup.

## REFERENCES

Wang, X., & Liu, Y. (2022). Machine Learning for Healthcare Data: A Study of *Random Forest* and *Naïve Bayes* in Predicting Health Risks. Journal of Medical Informatics, 18(2), 134-146. https://doi.org/10.1016/j.jmedinf.2022.03.005

Khan, S., & Tan, J. (2021). A Comparative Study of *Random Forest* and *Naïve Bayes* in Medical Risk Prediction. Computers in Biology and Medicine, 124(4), 1030-1042. https://doi.org/10.1016/j.compbiomed.2021.103653

Gupta, R., & Kumar, P. (2020). Comparing the Performance of *Random Forest* and *Naïve Bayes* Algorithms in Healthcare Predictions. Healthcare Analytics, 22(7), 289-301. https://doi.org/10.1016/j.healthanalytics.2020.09.017

Patel, R., & Shah, A. (2020). Predicting Hypertension Risk with Machine Learning Algorithms. Journal of Artificial Intelligence in Medicine, 45(2), 200-212. https://doi.org/10.1016/j.artmed.2020.02.009

Singh, D., & Prakash, M. (2021). Exploring the Effectiveness of Machine Learning Algorithms for Predicting Hypertension Risk. AI in Healthcare, 29(4), 113-125. https://doi.org/10.1109/aihc.2021.013022

Zhang, L., & Zhou, H. (2022). Performance Evaluation of *Random Forest* in Health Risk Classification: A Review. International Journal of Medical Informatics, 156(1), 12-25. https://doi.org/10.1016/j.ijmedinf.2021.104389

Kumar, A., & Sharma, S. (2021). A Comparative Study of Machine Learning Algorithms in Predicting Hypertension Risk in Large Datasets. Journal of Data Science and Machine Learning Applications, 12(3), 56-73. https://doi.org/10.1007/jdml.2021.00625

Sharma, R., & Gupta, A. (2020). *Random Forest* vs *Naïve Bayes*: A Comparative Approach to Predicting Hypertension Risk. Advances in Intelligent Systems and Computing, 930, 87-99. https://doi.org/10.1007/978-3-030-38076-2_9

Cheng, S., & Yu, X. (2020). Evaluation of Predictive Models for Hypertension Risk Using *Random Forest* and *Naïve Bayes*. Bioinformatics and Biology Insights, 14(1), 123-135. https://doi.org/10.1177/1177932219895425

Liu, Z., & Zhang, X. (2019). An Evaluation of *Random Forest* and *Naïve Bayes* Algorithms for Health Risk Prediction in Hypertension. Journal of Healthcare Informatics Research, 24(5), 450-462. https://doi.org/10.1007/s41155-019-0081-4