BAB 1
PENDAHULUAN

1.1 Latar Belakang

Hipertensi merupakan penyakit yang dijuluki sebagai "*the silent killer*" (Fatchanuraliyah dkk., 2024), karena menjadi salah satu penyakit yang cukup berbahaya. Hipertensi dapat menyerang siapa saja tanpa memberikan tanda apapun pada tubuh. Secara global, hipertensi memengaruhi sekitar 1,4 miliar orang dewasa yang berusia antara 30 hingga 79 tahun, dengan prevalensi yang terus meningkat seiring dengan pertumbuhan populasi serta perubahan gaya hidup yang terjadi (WHO, 2025). Survei Kesehatan Indonesia menunjukkan bahwa prevalensi hipertensi di Indonesia mencapai 10,7% di kalangan usia 18-24 tahun dan 17,4% pada kelompok usia 25-34 tahun (Badan Kebijakan Pembangunan Kesehatan, 2024). Kondisi tersebut menegaskan bahwa hipertensi tidak hanya menyerang kelompok usia lansia, tetapi juga mulai banyak dialami oleh kelompok usia produktif.

Hipertensi didefinisikan sebagai keadaan di mana tekanan darah sistolik seseorang setidaknya 140 mmHg atau lebih dan/atau tekanan darah diastoliknya minimal 90 mmHg (Purwono dkk., 2022). Prevalensi hipertensi berdasarkan hasil pengukuran pada kelompok usia 18 tahun ke atas mencapai 34,1%, tertinggi di daerah Kalimantan Selatan sebesar 44,1%, sedangkan terendah di daerah Papua sebesar 22,2% (Kemenkes, 2019). Melalui prevalensi hipertensi yang mencapai 34,1%, ternyata hanya sekitar 8,8% yang benar-benar terdiagnosis mengalami hipertensi. Salah satu faktor yang menjadi penyebab tingkat kontrol hipertensi masih rendah adalah kurangnya kesadaran masyarakat mengenai kondisi mereka sendiri, bahwa dirinya terkena hipertensi.

Hipertensi hingga kini menjadi salah satu masalah kesehatan utama di Indonesia, terutama karena kondisi ini sering dijumpai dalam layanan kesehatan masyarakat. Salah satu perkembangan teknologi di bidang kesehatan yang membantu para tenaga medis dalam mengklasifikasikan penyakit dengan mudah adalah melalui pemanfaatan *machine learning* (Telaumbanua dkk., 2019). *Machine learning* merupakan bagian dari kecerdasan buatan yang berfokus dalam

mengembangkan algoritma untuk mengolah dan memahami data dalam jumlah besar, khususnya untuk mengolah data klinis dan menemukan pola hubungan antaratribut kesehatan (Iksan dkk., 2025). Salah satu algoritma yang sering diterapkan dalam *machine learning* karena kemampuannya dalam menghasilkan prediksi yang akurat, seperti dalam hal pengenalan pola, analisis data biologis, serta prediksi tren pasar adalah algoritma *random forest*.

Terdapat beberapa penelitian terkait yang telah memanfaatkan *machine learning*, khususnya model klasifikasi *random forest* untuk menyelesaikan persoalan di bidang kesehatan. Algoritma *random forest* menghasilkan kinerja yang lebih baik dibandingkan dengan algoritma *k-nearest neighbor*, yakni tingkat akurasi sebesar 99,75%, dalam hal klasifikasi penyakit gagal ginjal (Salmon dkk., 2024). Penelitian lain menunjukkan tingkat akurasi hasil klasifikasi pada algoritma *support vector machine* adalah sebesar 94%, sedangkan untuk *random forest* menghasilkan akurasi lebih tinggi yakni sebesar 98% (Zam dkk., 2025). Penelitian selanjutnya, penggunaan algoritma *random forest* pada kasus klasifikasi gangguan tidur menghasilkan akurasi sebesar 89,69% pada kategori '*None*' dengan *recall* hanya sebesar 74,55% pada kategori '*Sleep Apnea*'. Sedangkan, untuk algoritma *K-Nearest Neighbor* (KNN) menghasilkan akurasi sebesar 87,02% dengan parameter k=5 (Khasanah dkk., 2025). Hasil tersebut menunjukkan bahwa algoritma *random forest* lebih unggul dibandingkan algoritma KNN, khususnya dalam klasifikasi kategori '*None*'.

Penelitian terkait telah menunjukkan keunggulan pemanfaatan *machine learning* pada model klasifikasi algoritma *random forest*, namun di sisi lain masih terdapat kendala dalam hal ketidakseimbangan kelas yang berdampak pada kemampuan algoritma dalam mengenali kategori dengan kelas minoritas. Penerapkan metode *Synthetic Minority Over-sampling Technique* (SMOTE) menjadi solusi untuk mengatasi masalah ketidakseimbangan kelas serta menghasilkan prediksi yang lebih akurat, di mana melalui model *neural network* yang dikombinasaikan dengan SMOTE mampu mencapai hasil akurasi sebesar 92%, presisi 91,88%, *recall* 92%, dan F1-*score* sebesar 91,91% (Misriati dan Aryanti, 2025). SMOTE pada dasarnya hanya diterapkan pada atribut numerik, sehingga ketika *dataset* terdapat kombinasi atribut numerik dan kategorikal,

diperlukan teknik SMOTE lain yang lebih sesuai yaitu *Synthetic Minority Over-sampling Technique-Nominal Continuous* (SMOTE-NC). Teknik *oversampling* SMOTE-NC pada algoritma *K-Nearest Neighbor* (KNN) mampu meningkatkan nilai sensitivitas secara signifikan pada konteks klasifikasi pasien gagal jantung, dengan nilai yang sebelumnya hanya 44,83% menjadi 72,41% (Priani dkk., 2025). Peningkatan ini membuktikan bahwa penggunaan metode *oversampling* bersama dengan algoritma *machine learning* efektif sebagai solusi masalah ketidakseimbangan kelas.

Penggunaan model klasifikasi *random forest* sering kali hanya diberikan satu nilai di setiap *hyperparameter*-nya, sehingga pengukuran peforma yang dihasilkan dapat tidak maksimal. Mengatasi masalah tersebut diperlukan proses *hyperparameter tuning*, salah satunya yaitu *GridSearchCV*. Terdapat penelitian terkait yang mengimplementasikan teknik optimasi *hyperparameter GridSearchCV* pada model klasifikasi *random forest* dalam memprediksi dini jenis hipertensi, yang di mana menggunakan data penelitian dari PPG-BP (*photoplethysmography-blood pressure*) sebanyak 219 sampel orang dewasa dengan rentang usia 21-86 tahun (Dewi dkk., 2022). Penelitian tersebut menunjukkan akurasi yang dihasilkan model sebelum dilakukan optimasi sebesar 72,3%. Setelah penerapan *GridSearchCV* menghasilkan parameter optimal untuk model *random forest*, yakni peningkatan akurasi hingga 86,1%. Berdasarkan temuan tersebut, menunjukkan bahwa optimasi *hyperparameter* mampu mengoptimalkan kinerja algoritma *machine learning*, yang dapat dipergunakan dalam proses diagnosa penyakit.

Berdasarkan latar belakang dan penelitian terkait, maka penelitian ini bertujuan untuk menerapkan penggunaan algoritma *random forest* untuk klasifikasi risiko penyakit hipertensi dengan memanfaatkan *Synthetic Minority Over-sampling Technique-Nominal Continuous* (SMOTE-NC) sebagai solusi untuk ketidakseimbangan kelas serta kombinasi *hyperparameter* yang mencakup parameter *n_estimators, max_depth*, dan *max_features*, agar kinerja model lebih optimal. Penelitian ini diharapkan dapat menjadi acuan dalam upaya pencegahan, serta mampu memberikan kontribusi di bidang kesehatan terutama bagi tenaga medis dalam melakukan diagnosis dini dan pencegahan penyakit hipertensi secara lebih efektif.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Seberapa besar peningkatan kinerja algoritma *random forest* dalam klasifikasi risiko penyakit hipertensi dengan penerapan *Synthetic Minority Over-sampling Technique-Nominal Continuous* (SMOTE-NC) jika dilihat dari metrik *accuracy*, *precision*, *recall*, dan F1-*score*?
2. Manakah kombinasi *hyperparameter* (*n_estimators*, *max_depth*, dan *max_features*) yang menghasilkan performa terbaik berdasarkan metrik *accuracy*, *precision*, *recall*, dan F1-*score* setelah penerapan SMOTE-NC?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, maka tujuan dalam penelitian ini adalah sebagai berikut:

1. Mengukur peningkatan performa algoritma *random forest* dalam mengklasifikasikan risiko penyakit hipertensi sebelum dan sesudah penerapan SMOTE-NC, guna membuktikan efektivitas penanganan *imbalanced class* terhadap metrik *accuracy*, *precision*, *recall*, dan F1-*score*.
2. Menganalisis kombinasi *hyperparameter* (*n_estimators, max_depth*, dan *max_features*) yang menghasilkan performa terbaik berdasarkan metrik *accuracy*, *precision*, *recall*, dan F1-*score* setelah penerapan SMOTE-NC.

1.4 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak, di antaranya:

1. Penelitian ini dapat digunakan sebagai wawasan ilmu pengetahuan dan bahan pembelajaran bagi mahasiswa terkait algoritma *random forest*.
2. Penelitian ini dapat memberikan informasi serta referensi terkait implementasi algoritma *random forest* dalam mengklasifikasikan risiko penyakit hipertensi.

1.5 Batasan Penelitian

Agar penelitian dapat terfokus dan mencapai tujuan yang telah ditetapkan, maka penelitian ini dibatasi pada hal-hal berikut:

1. Data yang digunakan dalam penelitian ini merupakan *dataset* publik yang diperoleh dari situs Kaggle. *Dataset* tersebut bernama *Hypertension-risk-model-main* yang dipublikasikan oleh Md Raihan Khan dan terakhir diperbarui pada tahun 2023. *Dataset* tersedia dalam format *Comma-Separated Values* (CSV) dan terdiri dari 4.240 *record* serta 13 atribut (kolom). Seluruh data dalam *dataset* digunakan sebagai dasar analisis untuk membangun model klasifikasi risiko penyakit hipertensi. *Dataset* diakses pada 28 Oktober 2025 melalui tautan: https://www.kaggle.com/datasets/khan1803115/hypertension-risk-model-main
2. Evaluasi performa model dilakukan menggunakan *confusion matrix*, dengan indikator meliputi *accuracy*, *precision*, *recall*, dan F1-*score*.
3. Algoritma yang digunakan dalam penelitian ini adalah *random forest*, tanpa dilakukan perbandingan secara langsung dengan algoritma lain.
4. Konfigurasi parameter utama dari *random forest* yang digunakan dalam penelitian ini adalah *n_estimators, max_depth*, dan *max_features*.