Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Klasifikasi Penyakit Hipertensi Menggunakan Metode *Random Forest*

Novianti[1], Syarifah Putri Agustini Alkadri[2*], Izhan Fakhruzi[3]
Teknik Informatika, Universitas Muhammadiyah Pontianak, Pontianak, Indonesia
*e-mail *Corresponding Author*: agustini.putri@unmuhpnk.ac.id

Abstract

This study discusses the classification of hypertension using the Random Forest method with a focus on age as the main factor. Given the serious impact of hypertension on health, research aims to simplify understanding of the problem, identify treatment gaps, and propose algorithm-based solutions. Using the PPG-BP Database, research methods involve problem identification, data collection, preprocessing, Random Forest modeling, hyperparameter tuning, and model evaluation. The findings show a high level of accuracy, 98% on training data and 95% on testing data, with the model being able to predict hypertension classification based on the variables age, blood pressure, heart rate and body mass index. Despite data imbalance, the preprocessing steps proved to be effective. The research conclusions contribute to the understanding of disease classification, especially hypertension, as well as practical guidance in efforts to prevent and treat it.

Keywords: *Classification; Data Mining; Hypertension; Random Forest*

Abstrak

Penelitian ini membahas klasifikasi penyakit hipertensi menggunakan metode Random Forest dengan fokus pada usia sebagai faktor utama. Dengan dampak serius hipertensi terhadap kesehatan, penelitian bertujuan untuk menyederhanakan pemahaman masalah, mengidentifikasi celah penanganan, dan mengusulkan solusi berbasis algoritma. Menggunakan PPG-BP Database, metode penelitian melibatkan identifikasi masalah, pengumpulan data, *preprocessing*, permodelan *Random Forest, tuning hyperparameter*, dan evaluasi model. Hasil temuan menunjukkan tingkat akurasi tinggi, 98% pada data training dan 95% pada data testing, dengan model mampu memprediksi klasifikasi hipertensi berdasarkan variabel usia, tekanan darah, detak jantung, dan indeks massa tubuh. Meskipun ada ketidakseimbangan data, langkah-langkah preprocessing terbukti efektif. Simpulan penelitian memberikan kontribusi pada pemahaman klasifikasi penyakit, khususnya hipertensi, serta panduan praktis dalam upaya pencegahan dan penanganannya.

Kata kunci: *Klasifikasi; Data Mining; Hypertension; Random Forest*

1. Pendahuluan

Hipertensi terjadi dalam kategori usia 31-44 tahun dengan persentase 31,6%, pada umur 45-54 tahun sebanyak 45,3% dan 55-64 tahun sebanyak 55,2% [1]. Hipertensi menjadi salah satu faktor penting penyebab beberapa penyakit seperti *stroke*, *gagal jantung*, *infark miokard*, *atrial fibrilasi*, penyakit *arteri perifer* hingga diseksi *aorta*. Sedangkan merujuk pada data yang diambil dari *Task Force Report on High Blood Pressure in Children and Adolescents* pada tahun 1987 dan 1996 mengemukakan beberapa definisi hipertensi, terdapat beberapa jenis tekanan darah yang terjadi pada manusia dewasa [2]. Tekanan darah normal, pra-hipertensi, stadium-1 hipertensi dan stadium-2 hipertensi. Tahapan pra-hipertensi biasanya tidak diikuti dengan gejala-gejala yang dirasakan oleh individu, namun berpotensi pada resiko menjadi penyakit hipertensi hingga penyakit kardiovaskular [3].

Data Kementerian Kesehatan Indonesia menggambarkan kondisi prevalensi hipertensi di berbagai kelompok usia. Namun, terdapat kesenjangan antara kondisi ril saat ini dengan kondisi ideal yang diharapkan dalam upaya pencegahan dan penanganan hipertensi. Kesenjangan ini menciptakan masalah yang dapat diukur dan perlu dicari solusi yang efektif.

Masalah klasifikasi sering dijumpai pada kehidupan nyata, pada masa salah satunya masalah Hipertensi adalah faktor penyebab timbulnya penyakit berat seperti *serangan jantung*, *gagal ginjal dan Stroke*. Tekanan darah orang dikatakan hipertensi apabila 140/90 mmHg dan 139/89 mmHg disebut prahipertensi sedangkan tekanan darah normal 120/80 mmHg. Terdapat berbagai penyakit *Stroke, jantung* dan *gagal ginjal*. *Stroke* merupakan salah satu penyakit yang menyebabkan kematian dan cacat tertinggi [4].

Analisis *Random Forest* adalah algoritma untuk *supervised learning* yang bisa digunakan untuk klasifikasi maupun regresi. Algoritma ini paling fleksibel dan mudah digunakan. Kelebihan *Random Forest* dapat mengatasi noise dan missing value serta mengatasi data yang besar, kekurangan random forest yaitu interpretasi yang sulit dan membutuhkan tuning model yang tepat untuk data. *Random Forest* merupakan salah satu algoritma pohon keputusan dari klasifikasi dengan tingkat akurasi yang baik [5][6]. *Random Forest* merupakan sebuah metode ensemble yang terdiri dari beberapa pohon keputusan sebagai *classifier*. Kelas yang dihasilkan dari proses klasifikasi ini diambil dari kelas terbanyak yang dihasilkan oleh pohon-pohon keputusan yang ada pada *Random Forest*. Dengan melakukan *voting* pada pohon-pohon keputusan yang tersedia membuat akurasi dari *Random Forest* meningkat [7].

Tujuan dari penulisan ini adalah untuk menyederhanakan pemahaman tentang masalah hipertensi, mengidentifikasi gap dalam penanganannya, dan mengusulkan solusi berbasis algoritma *Random Forest*. Manfaatnya diharapkan dapat memberikan kontribusi pada pemahaman lebih lanjut tentang klasifikasi penyakit, khususnya hipertensi, serta memberikan panduan praktis dalam upaya pencegahan dan penanganannya.

2. Tinjauan Pustaka

Penelitian sebelumnya oleh Kharits Abdul Khalim [8] tentang hipertensi menunjukkan bahwa hipertensi adalah kondisi tekanan darah tinggi. Penelitian tersebut bertujuan untuk memahami kinerja algoritma *Random Forest* dan *Naïve Bayes* dalam memprediksi penyakit hipertensi dengan menggunakan data seperti usia, jenis kelamin, tekanan darah, kolesterol, dan lainnya. Hasilnya menunjukkan bahwa *Random Forest* memiliki kinerja lebih baik dibandingkan dengan *Naïve Bayes*.

Kemudian Penelitian selanjutnya oleh M. Fahrul Rizki Aditya [9] menyoroti pentingnya isu kesehatan masyarakat terkait hipertensi, yang sering kali tidak menunjukkan gejala nyata pada pasien. Hipertensi diidentifikasi sebagai faktor risiko utama pada penyakit jantung koroner, gagal jantung, dan stroke. Dalam konteks perkembangan teknologi, penelitian ini berfokus pada penggunaan kecerdasan buatan, khususnya *decision tree* dan *random forest*, untuk mendeteksi penyakit hipertensi dengan tingkat akurasi optimal. Hasil penelitian menunjukkan bahwa kedua metode tersebut mampu mencapai akurasi 100%, memperkuat kontribusi mereka dalam mendiagnosis dan mengelola penyakit hipertensi secara efektif, memberikan dukungan potensial bagi profesional medis, terutama dokter dan rumah sakit.

Penelitian sebelumnya menunjukkan bahwa metode *Random Forest* memiliki kinerja yang baik dalam mengklasifikasi penyakit hipertensi. Namun, penelitian ini mengambil langkah lebih lanjut dengan fokus pada klasifikasi tahap-tahap dalam hipertensi, seperti normal, *Prehypertension*, stage 1, dan stage 2. Tujuan penelitian ini adalah untuk mendapatkan pemahaman yang lebih mendalam tentang kemampuan metode *Random Forest* dalam mengklasifikasi tahap-tahap hipertensi berdasarkan variabel-variabel seperti usia, jenis kelamin, tekanan darah, kolesterol, dan faktor-faktor lainnya. Dengan menggali lebih dalam pada tahap-tahap hipertensi, penelitian ini berusaha memberikan kontribusi yang lebih spesifik terhadap pemahaman dan manajemen penyakit ini.

3. Metodologi

Metodelogi Penelitian ini adalah jenis penelitian kuantitatif dengan menerapkan *teknik data mining* yang menggunakan metode *random forest*. Tahapan penelitian yang digunakan yaitu, Identifikasi Masalah, Pengumpulan Data, Pemrosesan *Data Mining*, Evaluasi Model Klasifikasi, Selesai.

3.1. Identifikasi Masalah

Permasalahan dalam penelitian ini adalah bagaimana cara mengatasi sulitnya mendiagknosa membuat klasifikasi pohon algoritma penyakit hipertensi dengan menggunakan *Random Forest*, dan bagaimana cara mengetahui nilai akurasi dari proses klasifikasi metode *Random Forest*.

3.2. Pengumpulan Data

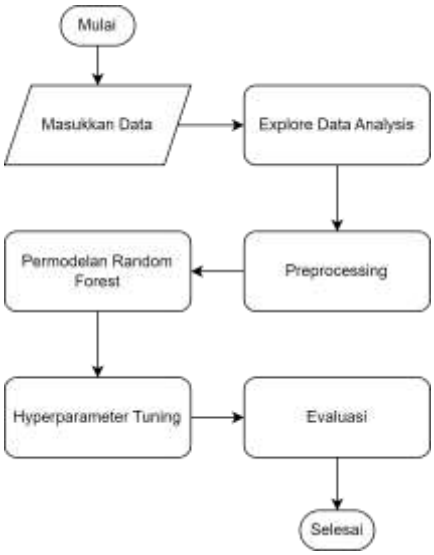
Pada tahapan ini bertujuan untuk mendapatkan data yang baik dengan mencari dataset yang sesuai dengan penelitian karena sulitnya mendapatkan data dari Indonesia karena kode etik medis yang mengharuskan kerahasiaan pasien jadi penelitian ini menggunakan dataset dari PPG-BP Database [10].

Table 1. Fitur dan Keterangan

No	Fitur	Keterangan	Jenis Data
1	Sex	Jenis kelamin pasien	Kategorikal
2	Age	Usia pasien	Numerikal
3	Height	Tinggi pasien	Numerikal
4	Wight	Berat badan pasien	Numerikal
5	Systolic Blood Pressure (SBP)	Tekanan darah sistolik	Numerikal
6	Diastolic Blood Pressure (DBP)	Tekanan darah diastolic	Numerikal
7	Heart Rate	Tekanan darah pasien	Numerikal
8	BMI	Indeks masa tubuh	Numerikal
9	Label	Label dataset	Kategorikal

3.3. Pemrosesan Data Mining

Pada proses ini dijelaskan alur penelitian dan langkah-langkah mengolah data agar hasilnya dapat berfungsi dengan baik dalam memprediksi penyakit hipertensi.



Gambar 1. Pemrosesan Data Mining

Pada diagram alir pada Gambar 1 menjelaskan alur atau Langkah-langkah dalam melakukan permodelan *Random Forest*, Adapun penjelasan lengkapanya adalah:

1) Input Data: Tahapan awal melibatkan pengumpulan dan pemasukan data ke dalam sistem menggunakan *library* atau modul khusus, seperti Pandas pada Python [11].
2) *Explore Data Analys*: Setelah input data, dilakukan analisis eksploratif data untuk memahami karakteristik, distribusi, dan pola data. EDA membantu dalam menentukan langkah-langkah selanjutnya dalam proses analisis.
3) *Preprocessing*: melibatkan langkah-langkah seperti Penanganan data hilang, Mengisi atau menghapus nilai yang hilang, Penanganan *outlier*, Deteksi dan penanganan data yang signifikan, Transformasi data, dan Melakukan *encoding* atau normalisasi jika diperlukan untuk persiapan data yang sesuai dengan kebutuhan model.
4) Permodelan *Random Forest*: Data yang telah melewati tahap preprocessing digunakan untuk melatih model klasifikasi menggunakan algoritma *Random Forest*. Model ini

dapat menghasilkan prediksi berdasarkan fitur-fitur yang telah dipilih. Dalam *Random Forest*, prediksi akhir didasarkan pada mayoritas suara dari semua pohon dalam ensemble. Dalam contoh ini, setiap pohon memberikan suara berdasarkan fitur-fitur yang dievaluasi, dan prediksi akhir *Random Forest* akan bergantung pada mayoritas suara dari pohon-pohon tersebut [12].

a. Menentukan data awal, Setiap pengambil keputusan memberikan nilai sesuai preferensinya yang menunjukkan kepentingan suatu kriteria tertentu.
b. Normalisasi data awal, Kurangkan tiap nilai kriteria dengan nilai paling ideal, hasil pengurangan tersebut dinyatakan k_{ij}
c. Menentukan nilai matriks (a_{ij})

$$a_{ij} = \frac{k_{ij}}{\sum_{i=1}^{m}\sum_{i=1}^{n} k_{ij}} \dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

a_{ij} : hasil perhitungan matriks data kriteria
k_{ij} : nilai setiap kriteria dari normalisasi data awal
i : responden ke 1,2,…. i
j : kriteria ke 1,2,… j
m : jumlah pengambil keputusan
n : jumlah kriteria

d. Perhitungan nilai entropy untuk setiap criteria

$$E_j = \left[\frac{-1}{ln\ m}\right]\sum_{i=1}^{n}\left[a_{ij}\ln(a_{ij})\right] \dots\dots\dots\dots\dots (2)$$

E_j : nilai bobot entropy
ln : nilai log dari total pengambil keputusan

e. Perhitungan dispersi untuk setiap kriteria

$$D_j = 1 = E_j \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

D_j : niali dispersi entropy.

f. Noramlisais nilai dispersi

$$W_j = \frac{D_j}{\sum D_j} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4)$$

w_j : nilai normalisasi dispersi (bobot prioritas kriteria)

5) *Hyperparameter Tuning*, Setelah permodelan, dilakukan *tuning* pada *hyperparameter* model untuk meningkatkan kinerja dan akurasi. Proses tuning membantu menyesuaikan parameter-model agar sesuai dengan data yang digunakan.
6) Evaluasi, Evaluasi model dilakukan untuk mengukur kinerja model terhadap data yang tidak terlihat selama pelatihan. Metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-score dapat digunakan untuk mengevaluasi sejauh mana model dapat memprediksi dengan benar.

3.4. Evaluasi

Evaluasi didapat berdasarkan dari analisis hasil pengujian yang telah dilakukan dan juga dari hasil analisis selama pembangunan aplikasi dan permodelan [13], pada penelitian ini dalakukan Eavaluasi mengguankan Matrix Evaluasi dan *Confusion Matrix* pada data pengujian dan pelatihan.

Table 2. Evaluasi Confusion Matrix

	1	2
1	True Positif	False Positif
2	False Negatif	True Negatif

Pada Tabel 2, ditampilkan isi dari Matriks Evaluasi Kesalahan (*Confusion Matrix*). Penjelasan dari *True Positive* (TP), *False Positive* (FP), *False Negative* (FN), dan *True Negative* (TN) adalah sebagai berikut [14] [15]:

1) *True Positive* (TP): Jumlah kasus yang benar-benar terdeteksi sebagai positif. Misalnya, model dengan benar memprediksi pasien hipertensi.
2) *False Positive* (FP): Jumlah kasus yang salah terdeteksi sebagai positif. Misalnya, model memprediksi pasien normal sebagai pasien hipertensi.
3) *False Negative* (FN): Jumlah kasus yang salah terdeteksi sebagai negatif. Contohnya, model memprediksi pasien hipertensi sebagai pasien normal.
4) *True Negative* (TN): Jumlah kasus yang benar-benar terdeteksi sebagai negatif. Misalnya, model dengan benar memprediksi pasien normal.

Dari hasil yang diambil pada matriks evaluasi kesalahan, kita dapat menghitung nilai akurasi, presisi, recall, dan skor F1.

1) Akurasi (Accuracy): Mengukur sejauh mana model memprediksi dengan benar, dihitung dengan (TP + TN) / (TP + TN + FP + FN).
2) Presisi (Precision): Mengukur tingkat ketepatan prediksi positif, dihitung dengan TP / (TP + FP).
3) Recall (Sensitivitas atau True Positive Rate): Mengukur kemampuan model untuk menemukan semua kasus positif, dihitung dengan TP / (TP + FN).
4) F1-Score: Mengukur keseimbangan antara presisi dan recall, dihitung dengan 2 * (Presisi * Recall) / (Presisi + Recall).

4. Hasil dan Pembahasan

Bab ini berisi uraian tentang hasil, analisis, dan pengujian aplikasi Implementasi metode random forest untuk klasifikasi kelangsungan hidup pasien hipetensis. Pengujian padapenelitian ini menggunakan tingkat akurasi dan pengujian antara metode, Tabel 3 adalah sampel data dari dataset yang digunakan dalam penelitian ini menggunakan dataset dari PPG-BP Database*:*

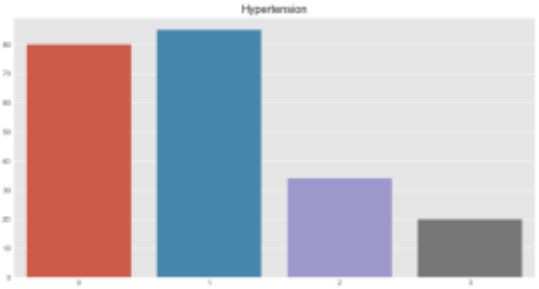
Table 3. Sampel Data

No	Sex	Age	Height	Weight	SBP	DBP	HeartRate	BMI	Label
1	F	45	152	63	161	89	97	27,27	Stage 2
2	F	50	157	50	160	93	76	20,28	Stage 2
3	F	47	150	47	101	71	79	20,89	Normal
4	M	45	172	65	136	93	87	21,97	Prehipertensi
5	F	46	155	65	123	73	73	27,06	Prehipertensi
6	F	48	160	68	124	62	70	26,56	Prehipertensi
7	F	48	153	49	126	78	84	20,93	Prehipertensi
8	F	53	160	70	108	73	84	27,34	Normal
9	M	60	169	71	153	72	85	24,86	Stage 1
10	F	47	150	47	98	56	69	20,89	Normal

Dari total 219 data, dilakukan pembagian data untuk pelatihan (latih) dan pengujian (uj) dengan proporsi 80% data latih dan 20% data uji. Pembagian ini dilakukan secara acak untuk memastikan representativitas kedua set data. Data latih digunakan untuk melatih model Random Forest, sementara data uji digunakan untuk menguji kinerja model dan mengukur tingkat akurasi prediksi.

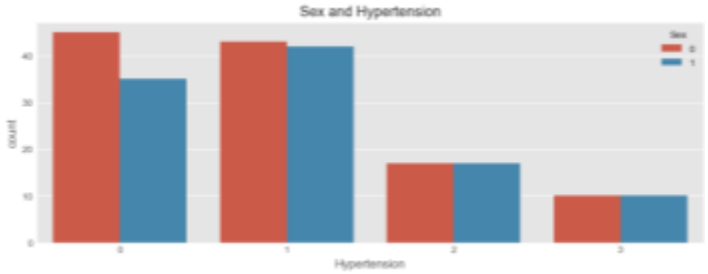
4.1. Hasil Explore Data Analysis (EDA)

Pada tahap ini dejelaskan hasil dari tahapan *explore data analysis*, hasil dari tahapan ini digunakan untuk mengetahui Langkah apa yang bisa diambil pada tahap berikutnya, dibawah ini adalah hasil-hasil yang didapat pada tahapan *explore data*.



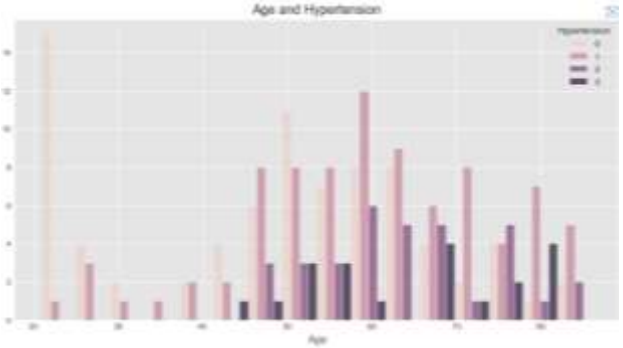
Gambar 2. EDA Label Dataset

Gambar 2 menggambarkan persebaran label pada dataset, yang terdiri dari empat target. Data terbanyak terdapat pada label nomor 1, dengan warna oren yang menandakan kondisi normal (pasien tidak terkena hipertensi). Selanjutnya, label biru menunjukkan prehipertensi, label ungu menandakan tahap 1 hipertensi, dan label hitam mengindikasikan tahap 2 hipertensi. Gambar tersebut memberikan representasi visual mengenai distribusi data pada fitur hipertensi dan label prediksi dalam aplikasi ini.



Gambar 3. EDA *Sex dan Hypertension*

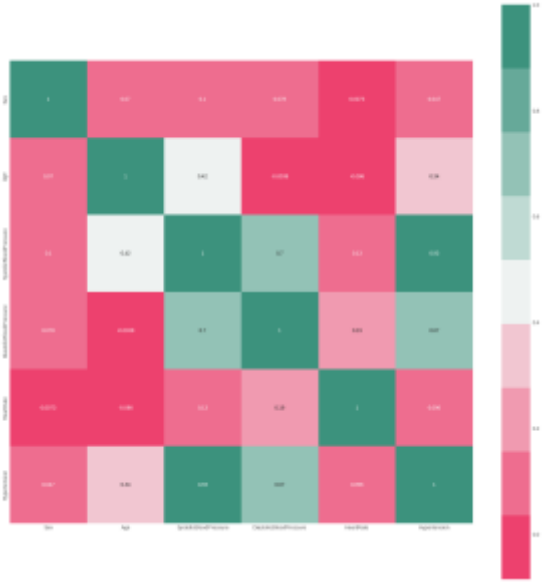
Gambar 3 memperlihatkan distribusi data hipertensi berdasarkan jenis kelamin, dan dapat disimpulkan bahwa lebih banyak pria yang terkena hipertensi dibandingkan wanita. Pada kategori normal (0), jumlah laki-laki lebih banyak, yaitu 48 dibandingkan dengan 35 perempuan. Sementara pada kategori prehipertensi (1), data EDA menunjukkan bahwa jumlah laki-laki (42) juga lebih banyak daripada perempuan (41). Tahap 1 (2) menunjukkan jumlah data laki-laki dan perempuan yang sama, yaitu 17, sedangkan pada tahap 2 (3) jumlahnya juga sama, yakni 10 untuk kedua jenis kelamin. Analisis ini memberikan gambaran mengenai distribusi hipertensi berdasarkan jenis kelamin dalam dataset tersebut.



Gambar 4. EDA *Age and Hypertension*

Gambar 4 memvisualisasikan sebaran hipertensi berdasarkan usia, dengan menunjukkan bahwa individu yang berusia di atas 40 tahun memiliki kecenderungan terkena hipertensi tahap 2. Kategori normal (0) menunjukkan data terbanyak untuk usia 20–30 tahun. Prehipertensi (1) paling banyak terjadi pada usia 40 tahun ke atas, dengan puncak terbanyak pada usia 60 tahun. Tahap 1 (2) hipertensi menunjukkan data terbanyak pada usia 60 tahun, sementara tahap 2 (3) hipertensi mencapai puncaknya pada rentang usia 70-80 tahun. Analisis ini

memberikan wawasan tentang hubungan antara usia dan tingkat keparahan hipertensi dalam dataset tersebut.



Gambar 5. EDA *Data Correlation*

Gambar 5 menggambarkan korelasi antar fitur, dan diperhatikan bahwa beberapa fitur menunjukkan tingkat korelasi yang beragam terhadap label y atau Hipertensi. Atribut Tekanan Darah Sistolik memiliki korelasi yang sangat baik dengan label, ditandai dengan nilai 0,93. Tekanan Darah Diastolik menunjukkan korelasi cukup baik dengan nilai 0,67. Sebaliknya, atribut Usia memiliki korelasi yang rendah dengan nilai 0,34, sementara atribut Jenis Kelamin (*Sex*) memiliki korelasi cukup baik dengan nilai sekitar 0,047. Detak Jantung (*Heart Rate*) juga menunjukkan korelasi baik dengan label, diperkirakan sekitar 0,096. Analisis korelasi ini memberikan gambaran tentang sejauh mana setiap fitur berkontribusi terhadap prediksi kondisi hipertensi, dengan korelasi tinggi menandakan pengaruh yang lebih besar.

4.2. Hasil Preprocessing

Pada tahapan ini dilakukan beberapa penanganan untuk mempersiapkan data agar dapat dilatih menggunkan metode *Random Fores*.

1) Hasil Penanganan Data Hilang (*Missing Value*)

Langkah awal dari tahapan ini adalah menangani *Missing Value* dengan cara mengubah data kosong menjadi median atau nilai rata-rata.

```
Num                       0
subject_ID                0
Sex                       0
Age                       0
Height                    0
Weight                    0
SystolicBloodPressure     2
DiastolicBloodPressure    1
HeartRate                 0
BMI                       0
Hypertension              0
dtype: int64
```

Gambar 6. Memeriksa Missing Value

Pada gambar 6 adalah pengecekan data hilang, dan didapatkan data hilang pada fitur *Systolic Blood Pressure* berjumlah 2, dan *Diastolic Blood Pressure* yang berjumlah 1, akan dilakukan penanganan dengan merubah nilai hilang menjadi nilai *median.*

2) Hasil Penanganan Data Kembar (*Duplicate*)

Langakah selanjutnya yaitu menangani data *duplicate* dengan menghapus data tersebut untuk mencegah terjadinya *outlier*, akan tetapi setelah dilakukan pengecekan tidak didapat juga data *duplicate* didalam dataset ini, jadi penulis tidak melakukan penanganan apapun.

3) Hasil Penanganan *Label Encoding Data*

Tujuan dari tahap ini yaitu merubah tipe data yang berupa objek menjadi integer agar mesin dapat membaca data tersebut semisal jenis kelamin Peria menjadi 0 dan Wanita menjadi 1.

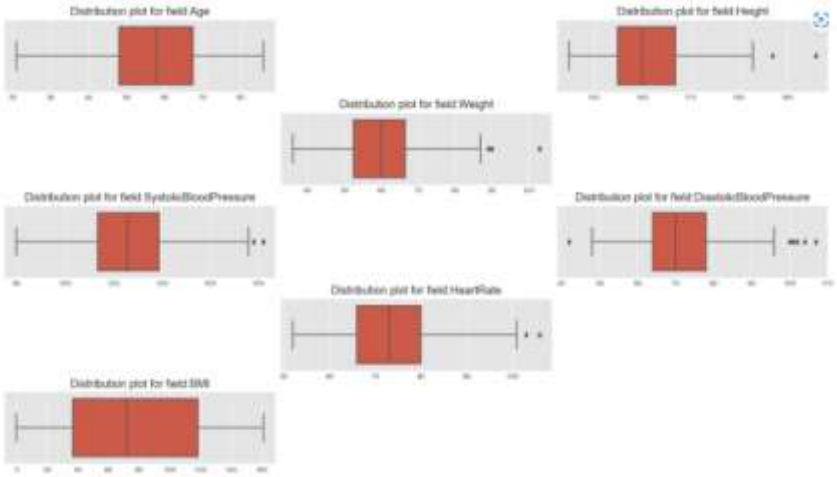
```
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Num                    219 non-null    int64
 1   subject_ID             219 non-null    int64
 2   Sex                    219 non-null    object
 3   Age                    219 non-null    int64
 4   Height                 219 non-null    int64
 5   Weight                 219 non-null    int64
 6   SystolicBloodPressure  219 non-null    int64
 7   DiastolicBloodPressure 219 non-null    int64
 8   HeartRate              219 non-null    int64
 9   BMI                    219 non-null    object
 10  Hypertension           219 non-null    object
dtypes: int64(8), object(3)
```

Gambar 7. Memeriksa Data Objek

Pada gamabr 7 adalah gambar hasil pengecekan tipe data dan didapati data objek pada fiture *sex*, BMI, dan *Hypertension*, pada penelitian ini pe nulis melakukan perubahan objek kedalam integer menggunakan *liblary python* yaitu *Label Encoder*.

4) Hasil Penanganan *Outlier*

Outlier adalah saat diaman mesin mempelajari keseluruhan data meskipun data itu salah oleh karena itu dilakukan penanganan dengan mnghapus data *outlier* atau merubah nilai data tersebut agar tidak berbanding sangat jauh dengan data-data yang lainya.

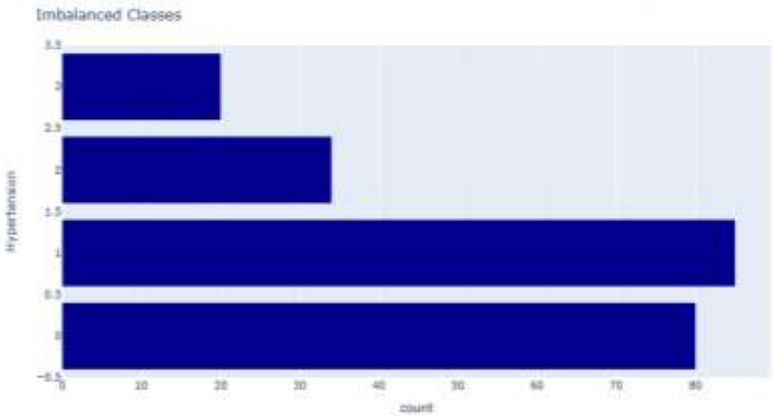


Gambar 8. Hasil Memeriksa Outlier

Pada gambar 8 adalah hasil pengecekan *outlier*, dan didapatkan hasil bahwa data tidak seimbang tidak begitu parah sehingga penulis tidak melakukan penanganan apapun karena ditakutkan merubah nilai asli dari data tersebut sehingga menurunkan tingkat akurasi dari permodelan *Random Forest*.

5) Hasil Penanganan Data Tidak Seimbang

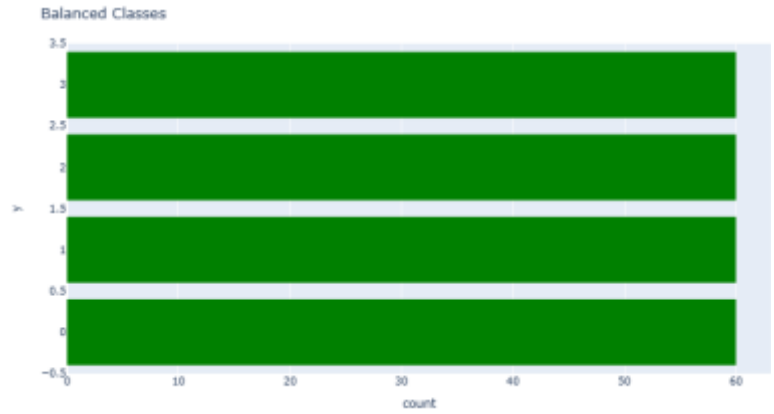
Langkah selanjutnya yitu menangani data tidak seimbang agar pelahitihan metode *Random Forest* menjadi lebih maksimal.



Gambar 9. Hasil memeriksa data balancing

Pada gambar 9 diatas adalah hasil pengecekan data tidak seimbang, dan dapat dilihat bahwa data diatas memiliki ketidak seimbangan sangat jauh dari nilai 0 hingga 3 ada yang berjumlah 80 pasien dan ada yang 20 pasien, Adapun penjelasanya sebagi berikut.

1) 0 (*Normal*): Pada gambar EDA diatas menunjukkan data Normal berjumlah sekitar 80 pasien.
2) 1 (*Prehypertension*): Pada gambar EDA diatas menunjukkan data Persiapan berjumlah sekitar 85 pasien.
3) 2 (*Stage 1*): Pada gambar EDA diatas menunjukkan data *Stage 1* berjumlah sekitar 34 pasien
4) 3 (*Stage 2*): Pada gambar EDA diatas menunjukkan data *Stage 2* berjumlah sekitar 20 pasien.



Gambar 10. Data setelah balancing.

Pada gambar 10 adalah gambar hasil setelah data dinormalisasikan menggunkan *liblary Python* yaitu SMOTETomek, Metode SMOTE (*Synthetic Minority Over-sampling Technique*) dan Tomek adalah dua metode yang dapat digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset.

Metode SMOTE dan Tomek dapat digunakan bersama-sama sebagai pendekatan gabungan untuk menyeimbangkan dataset. Pertama, metode Tomek dapat diterapkan untuk membersihkan data dari pasangan Tomek yang mungkin ada. Kemudian, metode SMOTE dapat digunakan untuk melakukan *oversampling* pada kelas minoritas yang tersisa. Pendekatan ini dapat membantu meningkatkan keseimbangan dataset dan kualitas pemodelan pada masalah klasifikasi dengan ketidakseimbangan kelas, dengan menggabungkan dua metode diatas didapatkan hasil dari mengatasi ketidakseimbangan kelas menjadi seniali 60 pada setiap fitur.

6) Hasil Penanganan Normalisasi

Pada tahapan ini dilakukan Normalisasi data atau pensekalaan data dari -1 hingga 1 dengan melakukan hal tersebut akan mempermudah mesin untuk menghitung, pada penelitian ini digunakan *MinMaxScaler* untuk menormalisasikan data, akan tetapi data akan dirubah Kembali menjadi nilai awal untuk aplikasi streamlit.

7) Hasil Fiture Selection

Dari hasil fitur selection ini, terdapat beberapa fitur yang terpilih, yaitu: *Age*, *SystolicBloodPressure*, *DiastolicBloodPressure*, *HeartRate*, dan BMI.

Pemilihan fitur-fitur ini menunjukkan bahwa dalam analisis data atau pembuatan model, fitur-fitur ini dianggap paling penting atau memiliki kontribusi signifikan terhadap hasil yang diinginkan. Misalnya, usia (*Age*) dapat menjadi indikator penting dalam beberapa analisis yang terkait dengan faktor usia, seperti risiko penyakit atau prediksi hasil tertentu. Tekanan darah sistolik (*SystolicBloodPressure*) dan tekanan darah diastolik (*DiastolicBloodPressure*) umumnya digunakan dalam evaluasi kesehatan dan pemantauan tekanan darah. *HeartRate* (detak jantung) juga merupakan parameter penting dalam banyak analisis yang berhubungan dengan kondisi kardiovaskular. BMI (*Body Mass Index*) adalah indikator yang digunakan untuk mengevaluasi kelebihan berat badan atau obesitas pada seseorang, Dengan menggunakan fitur-fitur ini, kita dapat meningkatkan kinerja model atau analisis data yang berkaitan dengan topik-topik tersebut, serta mengurangi kompleksitas dataset dengan memilih fitur-fitur yang paling informatif dan relevan.

4.2. Hasil Pelatihan

Hasil pengujian akurasi menggunakan permodelan Random Forest Entropy mendapatkan hasil seperti gambar 12

Training Entropy
98.828125

	precision	recall	f1-score	support
0	0.98	0.97	0.98	62
1	0.97	0.98	0.98	61
2	1.00	1.00	1.00	66
3	1.00	1.00	1.00	67
accuracy			0.99	256
macro avg	0.99	0.99	0.99	256
weighted avg	0.99	0.99	0.99	256

Testing Entropy
95.45454545454545

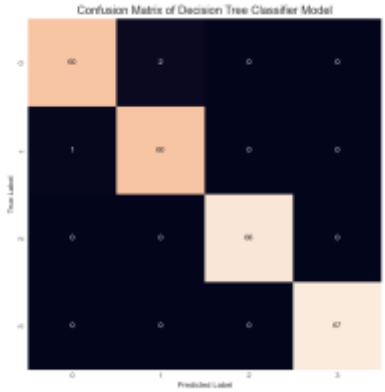
	precision	recall	f1-score	support
0	1.00	0.94	0.97	17
1	0.90	1.00	0.95	18
2	1.00	0.83	0.91	6
3	1.00	1.00	1.00	3
accuracy			0.95	44
macro avg	0.97	0.94	0.96	44
weighted avg	0.96	0.95	0.95	44

Gambar 11. Hasil Pengujian Matrik Evaluasi

Pada gambar 11, terdapat *matriks evaluasi* model perhitungan *Entropy*. Hasilnya menunjukkan bahwa model mencapai tingkat akurasi sebesar 98% pada data *training* dan 95% pada data *testing*. Akurasi sebesar 98% pada data *training* berarti model berhasil memprediksi dengan benar sekitar 98% dari data yang digunakan saat melatih model. Sedangkan, akurasi sebesar 95% pada data *testing* menunjukkan bahwa model mampu melakukan prediksi dengan tingkat keakuratan sekitar 95% pada data yang tidak digunakan saat melatih model, sehingga dapat menggeneralisasi dengan baik pada data baru.

1) Hasil *Matriks Evaluasi Training* menunjukkan bahwa model telah mencapai hasil yang sempurna dalam memprediksi label 2 dan 3. Ini berarti model berhasil mengklasifikasikan semua data dengan label 2 dan 3 dengan benar. Namun, untuk label 0 dan 1, meskipun tidak sempurna, model masih memberikan hasil yang cukup baik dalam memprediksi dengan akurasi yang tinggi.

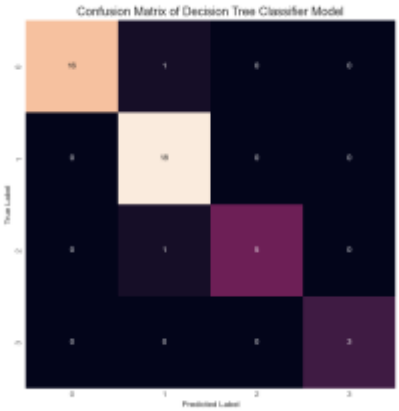
2) Hasil *Matriks Evaluasi Testing* menunjukkan bahwa model telah mencapai hasil yang sempurna dalam memprediksi label 0 dan 3. Ini berarti model berhasil mengklasifikasikan semua data dengan label 0 dan 3 dengan benar. Namun, untuk label 1 dan 2, meskipun tidak sempurna, model masih memberikan hasil yang cukup baik dalam memprediksi dengan akurasi yang tinggi.



Gambar 12. Confusion Matrix data Training

Hasil Dari *Confusion Matrix* perhitungan *entropy* pada data *Training* dengan akurasi 98% sebagai berikut:

1) Dari 60 data test untuk kelas Normal, sistem tersebut memprediksi dengan benar 60 data sebagai Normal, 2 data sebagai Prehypertension, 0 data sebagai Stage 1 hypertension, dan 0 data sebagai Stage 2 hypertension.
2) Dari 60 data test untuk kelas Prehypertension, sistem tersebut memprediksi dengan benar 60 data sebagai Prehypertension, 0 data sebagai Normal, 1 data sebagai Stage 1 hypertension, dan 0 data sebagai Stage 2 hypertension.
3) Dari 66 data test untuk kelas Stage 1 hypertension, sistem tersebut memprediksi dengan benar 66 data sebagai Stage 1 hypertension, 0 data sebagai Normal, 0 data sebagai Prehypertension, dan 0 data sebagai Stage 2 hypertension.
4) Dari 67 data test untuk kelas Stage 2 hypertension, sistem tersebut memprediksi dengan benar 67 data sebagai Stage 2 hypertension, 0 data sebagai Normal, 0 data sebagai Prehypertension, dan 0 data sebagai Stage 1 hypertension.



Gambar 13. Confusion Matrix data Testing

Hasil Dari *Confusion Matrix* perhitungan *entropy* pada data Testing dengan akurasi 95% pada gambar 5.13 adalah sebagai berikut:

1) Dari 16 data test untuk kelas Normal, sistem tersebut memprediksi dengan benar 16 data sebagai Normal, 1 data sebagai Prehypertension, 0 data sebagai Stage 1 hypertension, dan 0 data sebagai Stage 2 hypertension.
2) Dari 18 data test untuk kelas Prehypertension, sistem tersebut memprediksi dengan benar 18 data sebagai Prehypertension, 0 data sebagai Normal, 0 data sebagai Stage 1 hypertension, dan 0 data sebagai Stage 2 hypertension.
3) Dari 5 data test untuk kelas Stage 1 hypertension, sistem tersebut memprediksi dengan benar 5 data sebagai Stage 1 hypertension, 1 data sebagai Normal, 0 data sebagai Prehypertension, dan 0 data sebagai Stage 2 hypertension.
4) Dari 3 data test untuk kelas Stage 2 hypertension, sistem tersebut memprediksi dengan benar 3 data sebagai Stage 2 hypertension, 0 data sebagai Normal, 0 data sebagai Prehypertension, dan 0 data sebagai Stage 1 hypertension.

4.3. Pembahasan

Hasil penelitian ini menunjukkan bahwa implementasi metode Random Forest pada klasifikasi hipertensi berhasil mencapai tingkat akurasi yang tinggi, dengan 98% pada data training dan 95% pada data testing. Temuan dari EDA konsisten dengan literatur terdahulu, mengidentifikasi faktor risiko seperti jenis kelamin, usia, dan tekanan darah. Meskipun terdapat ketidakseimbangan data, langkah-langkah preprocessing, termasuk SMOTETomek, efektif dalam menangani masalah ini. Model Random Forest mampu mengklasifikasikan kelangsungan hidup pasien hipertensi dengan baik. Hasil ini tidak hanya mendukung temuan penelitian terdahulu tetapi juga menambah nilai dengan penerapan metode baru dalam konteks kesehatan hipertensi. Saran untuk penelitian selanjutnya adalah melibatkan dataset yang lebih besar dan variabel yang lebih komprehensif untuk meningkatkan generalisasi model.

5. Simpulan

Berdasarkan penelitian ini, dapat disimpulkan bahwa metode Random Forest Classification efektif dalam mengklasifikasikan penyakit hipertensi berdasarkan usia, dengan tingkat akurasi yang baik. Model klasifikasi yang dikembangkan mampu memprediksi klasifikasi hipertensi dengan akurasi sebesar 98% pada data training dan 95% pada data testing. Temuan ini menunjukkan bahwa metode Random Forest memiliki potensi yang besar dalam mengatasi masalah klasifikasi kompleks, terutama dalam hal mengidentifikasi tingkat risiko hipertensi berdasarkan usia. Selain itu, penelitian ini menunjukkan bahwa fitur-fitur seperti usia, tekanan darah sistolik, tekanan darah diastolik, detak jantung, dan indeks massa tubuh (BMI) memiliki kontribusi signifikan dalam mengklasifikasikan hipertensi. Rekomendasi untuk penelitian mendatang adalah mempertimbangkan pengembangan model dengan memperluas cakupan fitur dan meningkatkan keseimbangan data untuk meningkatkan ketepatan dalam mengklasifikasikan penyakit hipertensi. Selain itu, diperlukan validasi lebih lanjut terhadap model ini dengan dataset yang lebih luas dan perlu dipertimbangkan juga penerapan model ini dalam pengembangan sistem prediksi risiko hipertensi secara real-time untuk membantu pencegahan penyakit yang lebih efektif.

Daftar Referensi

[1] H. J, J. Andri, T. D. Payana, M. B. Andrianto, and A. Sartika, “Kualitas Tidur Berhubungan dengan Perubahan Tekanan Darah pada Lansia,” *J. Kesmas Asclepius*, vol. 2, no. 1, pp. 1–11, 2020, doi: 10.31539/jka.v2i1.1146.

[2] S. Hanny and Setyani2 Hanny, “Jurnal kedokteran dan kesehatan,” *Kefir a new role as nutraceuticals*, vol. 7, no. 5, pp. 200–209, 2016.

[3] P. Purwono *et al.*, “Model Prediksi Otomatis Jenis Penyakit Hipertensi Dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network,” *Insect (Informatics Secur. J. Tek. Inform.*, vol. 7, no. 2, pp. 82–90, 2022.

[4] J. H. Saing, “Hipertensi pada Remaja,” *Sari Pediatr.*, vol. 6, no. 4, p. 159, 2016, doi: 10.14238/sp6.4.2005.159-65.

[5] S. Adi, “Komparasi Metode Support Vector Machine (Svm), K-Nearest Neighbors (Knn), Dan Random Forest (Rf) Untuk Prediksi Penyakit Gagal Jantung,” *J. Ilm. Mat.*, vol. 10, no. 02, pp. 258–268, 2022.

[6] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK J.*

Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 21, no. 3, pp. 677–690, 2022, doi: 10.30812/matrik.v21i3.1726.

[7] I. R. Hikmah and R. N. Yasa, "Perbandingan Hasil Prediksi Diagnosis pada Indian Liver Patient Dataset (ILPD) dengan Teknik Supervised Learning Menggunakan Software Orange," *J. Telemat.*, vol. 16, no. 2, pp. 69–76, 2021.

[8] K. Abdul Khalim, U. Hayati, and A. Bahtiar, "Perbandingan Prediksi Penyakit Hipertensi Menggunakan Metode Random Forest Dan Naïve Bayes," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 498–504, 2023, doi: 10.36040/jati.v7i1.6376.

[9] M. F. R. Aditya and N. L. Azizah, "Prediction of Hypertension Disease Using Decision Tree and Random Forest Methods [Prediksi Penyakit Hipertensi Menggunakan metode Decision Tree dan Random Forest]," pp. 1–9, doi: https://doi.org/10.21070/ups.3200.

[10] H. Malik, M. Nuh, and M. H. Fatoni, "Perancangan Database Informasi Medis untuk Sistem Prediksi Hipertensi," *J. Tek. ITS*, vol. 9, no. 1, 2020, doi: 10.12962/j23373539.v9i1.45686.

[11] P. L. Gilabert *et al.*, "An efficient combination of digital predistortion and ofdm clipping for power amplifiers," *Int. J. RF Microw. Comput. Eng.*, vol. 19, no. 5, pp. 583–591, 2009, doi: 10.1002/mmce.20381.

[12] J. R. Thompson and B. L. Licklider, "Visualizing Urban forestry: Using concept maps to assess student performance in a learning-centered classroom," *J. For.*, vol. 109, no. 7, pp. 402–408, 2011.

[13] I. N. Abrar and A. Abdullah, "Klasifikasi Penyakit Liver Menggunakan Metode Elbow Untuk Menentukan K Optimal pada Algoritma K-Nearest Neighbor (K-NN)," vol. 12, pp. 218–228, 2023.

[14] A. Harun and D. Putri Ananda, "Analisa Sentimen Opini Publik Tentang Vaksinasi Covid-19 di Indonesia Menggunakan Naïve bayes dan Decission Tree," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 58–64, 2021, doi: 10.57152/malcom.v1i1.63.

[15] T. Praningki and I. Budi, "Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN," *Creat. Inf. Technol. J.*, vol. 4, no. 2, p. 83, 2018, doi: 10.24076/citec.2017v4i2.100.