BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil

Berdasarkan metodologi yang telah dirancang pada kasus prediksi kelangsungan hidup pasien penderita *cardiovascular disease* dengan metode *Random Forest* didapatkan hasil sebagai berikut.

a. *Data Collecting*

Penelitian ini menggunakan dataset penyakit *cardiovascular* yang diperoleh dari repositori Kaggle. Dataset yang digunakan berjumlah 3855 data dalam bentuk *comma separated values* (csv).

Tabel 4. 1 Metadata

No	Fitur	Tipe Data	Keterangan	Rentang Nilai
1	Age	Continues	Usia Pasien	1-100 (Tahun)
2	Diastolic BP	Continues	Tekanan Darah Diastolik	Normal : < 80 mmHg Tinggi: 80 - 84 mmHg Tahap 1 Hipertensi: 85 - 89 mmHg Tahap 2 Hipertensi: > 90 mmHg
3	Race	Categorical	Ras	1 : Putih, 2: Negro
4	Red blood cells	Continues	Sel darah merah	Pria: 4.7 – 6.1 mikroliter (mcL) Wanita: 4.2 – 5.4 mikroliter (mcL)

5	Sedimentation rate	Continues	Tingkat sedimentasi	Normal: Pria Dewasa: < 15 mm/jam Wanita Dewasa: < 20 mm/jam Anak-anak: < 10 mm/jam
6	Serum Albumin	Continues	Serum albumin	Normal: 3.5 – 5.0 g/dL(desiliter) Rendah (Hipoalbuminemia): < 3.5 g/dL Tinggi (Hiperalbuminemia): > 5.0 g/dL
7	Serum Cholesterol	Continues	Serum Kolesterol	Kolesterol diinginkan.: < 200 mg/dL Kolesterol Jahat : < 100 mg/dL Kolesterol Baik : > 60 mg/dL
8	Serum Iron	Continues	Serum Besi	60 - 170 mcg/dL
9	Serum Magnesium	Continues	Serum Magnesium	Rendah : < 1.5 mEq /L Normal : 1.5 – 2.5 mEq /L (miliekuivalen per liter) Tinggi : >1.5 mEq /
10	Serum Protein	Continues	Protein Serum	60 - 80 g/L
11	Sex	Categorical	Jenis Kelamin	Pria : 1, Wanita : 2

12	Systolic BP	Continues	Tekanan Darah Sistolik	Normal: < 120 mmHg Tinggi: 120 - 129 mmHg Tahap 1 Hipertensi: 130 - 139 mmHg Tahap 2 Hipertensi: > 140 mmHg
13	TIBC	Continues	Kapasitas Pengikatan Besi Total	240 - 450 mcg/dL
14	TS	Continues	Saturasi Transferrin	20 - 50
15	White blood cells	Continues	Sel Darah Putih	Rendah : < 4 mcL Normal : 4 – 11 mcL Tinggi : > 11
16	BMI	Continues	Indeks Massa Tubuh	Kurang: < 18,5 kg/m² Sehat: 18.5 – 24.9 kg/m² Obesitas : 25.0 – 29.9 kg/m² Obesitas Sedang: 30 – 34.9 kg/m² Obes Berat Berat: 35 – 39.9 kg/m² Obesitas Sangat Parah : > 40 kg/m²
17	Pulse pressure	Continues	Tekanan Nadi	Rendah : < 40 mmHg Normal : 40-60 mmHg Tinggi : > 60 mmHg
18	Y	Continues	Time (Waktu Kelangsungan Hidup Pasien)	-

b. Exploratotory Data Analysis

Adapun *sample* data penyakit *cardiovascluar* pada *google colaboratory* sebagai berikut.

	Age	Diastolic BP	Race	Red blood cells	Sedimentation rate	Serum Albumin	Serum Cholesterol	Serum Iron	Serum Magnesium	Serum Protein	Sex	Systolic BP	TIBC	TS	White blood cells	BMI	Pulse pressure	y
0	35.0	92.0	2.0	77.7	12.0	5.0	165.0	135.0	1.3	76.0	2.0	142.0	323.0	41.8	5.8	31.0	50.0	15.274658
1	71.0	78.0	2.0	77.7	37.0	4.0	298.0	89.0	13.8	6.4	2.0	156.0	331.0	26.9	5.3	32.0	78.0	11.586073
2	74.0	86.0	2.0	77.7	31.0	38.0	222.0	115.0	1.3	7.4	2.0	170.0	299.0	38.5	8.1	25.0	84.0	8.149087
4	32.0	70.0	2.0	77.7	18.0	5.0	203.0	192.0	1.3	7.3	1.0	128.0	386.0	49.7	8.1	20.0	58.0	-0.000000
12	51.0	76.0	1.0	77.7	16.0	4.4	182.0	116.0	1.6	8.0	2.0	148.0	331.0	35.0	10.4	17.0	72.0	20.053196

Gambar 4. 1 sample data

1) Missing Value

Pada proses *exploratory data analysis* perlu dilakukan pengecekan *missing value* terlebih dahulu pada setiap variabelnya.

Age	0
Diastolic BP	28
Race	0
Red blood cells	0
Sedimentation rate	357
Serum Albumin	0
Serum Cholesterol	0
Serum Iron	0
Serum Magnesium	0
Serum Protein	0
Sex	0
Systolic BP	27
TIBC	0
TS	0
White blood cells	447
BMI	0
Pulse pressure	28
y	0

Gambar 4. 2 missing value

Setelah dilakukan pengecekan *missing value* terdapat beberapa missing values pada variabel-variabel *Diastolic* BP, *Sedimentation rate, Systolic BP, White blood cells,* dan *Pulse pressure.* Untuk itu perlu dilakukan penanganan *missing value* terhadap variabel-variabel tersebut.

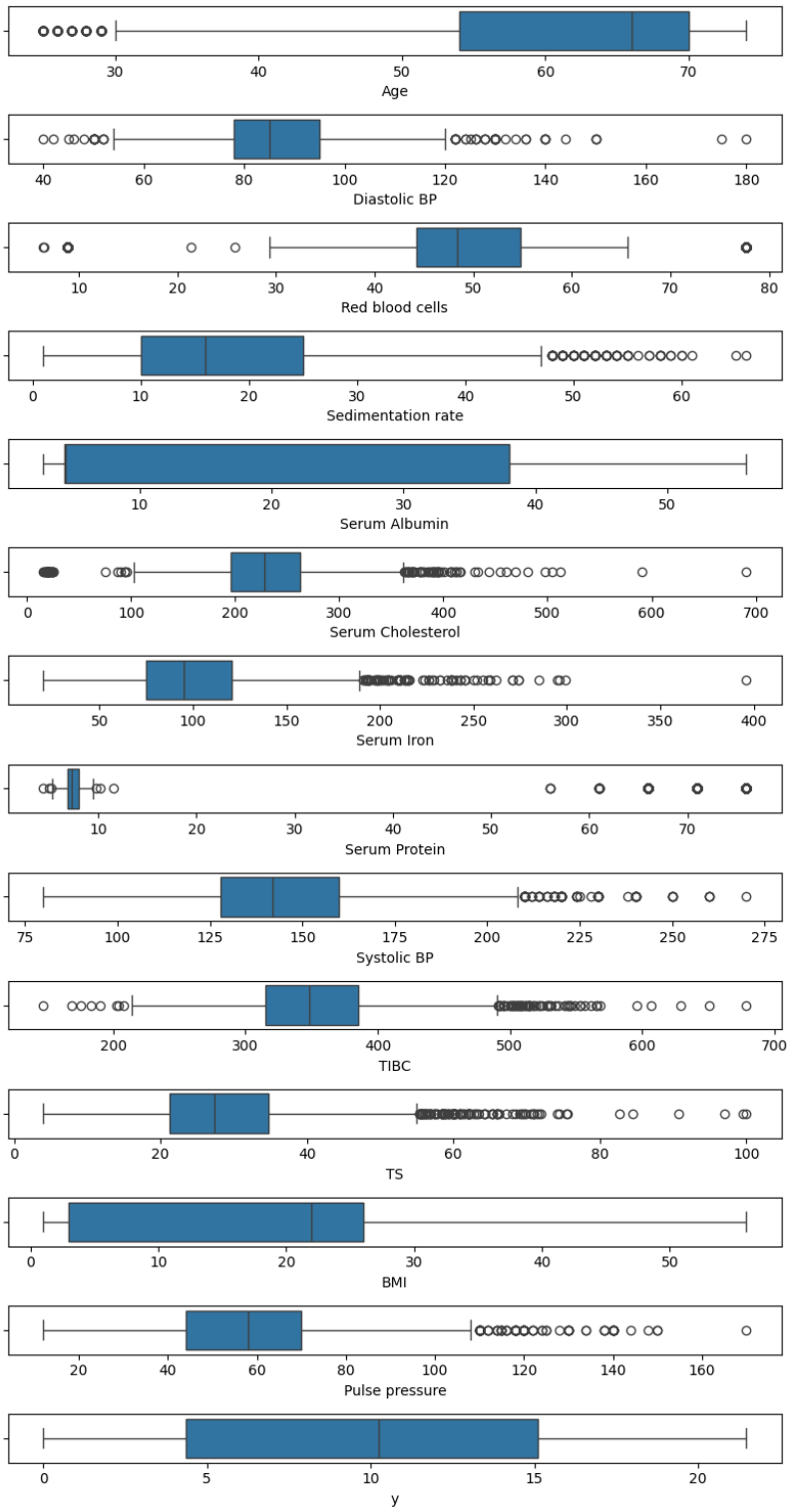
Age	0
Diastolic BP	0
Race	0
Red blood cells	0
Sedimentation rate	0
Serum Albumin	0
Serum Cholesterol	0
Serum Iron	0
Serum Magnesium	0
Serum Protein	0
Sex	0
Systolic BP	0
TIBC	0
TS	0
White blood cells	0
BMI	0
Pulse pressure	0
y	0

Gambar 4. 3 mengatasi missing value

Karena banyaknya jumlah *missing values* pada variabel *sedimentation rate* dan *white blood cells.* Sehingga *missing values* diatasi menggunakan metode imputasi *median*. Dengan metode ini, semua *missing values* telah berhasil dihilangkan dan siap untuk analisis lebih lanjut.

2) Outlier

Pada proses EDA selanjutnya dilakukan pengecekan outlier untuk masing-masing variabel dalam *dataset.*

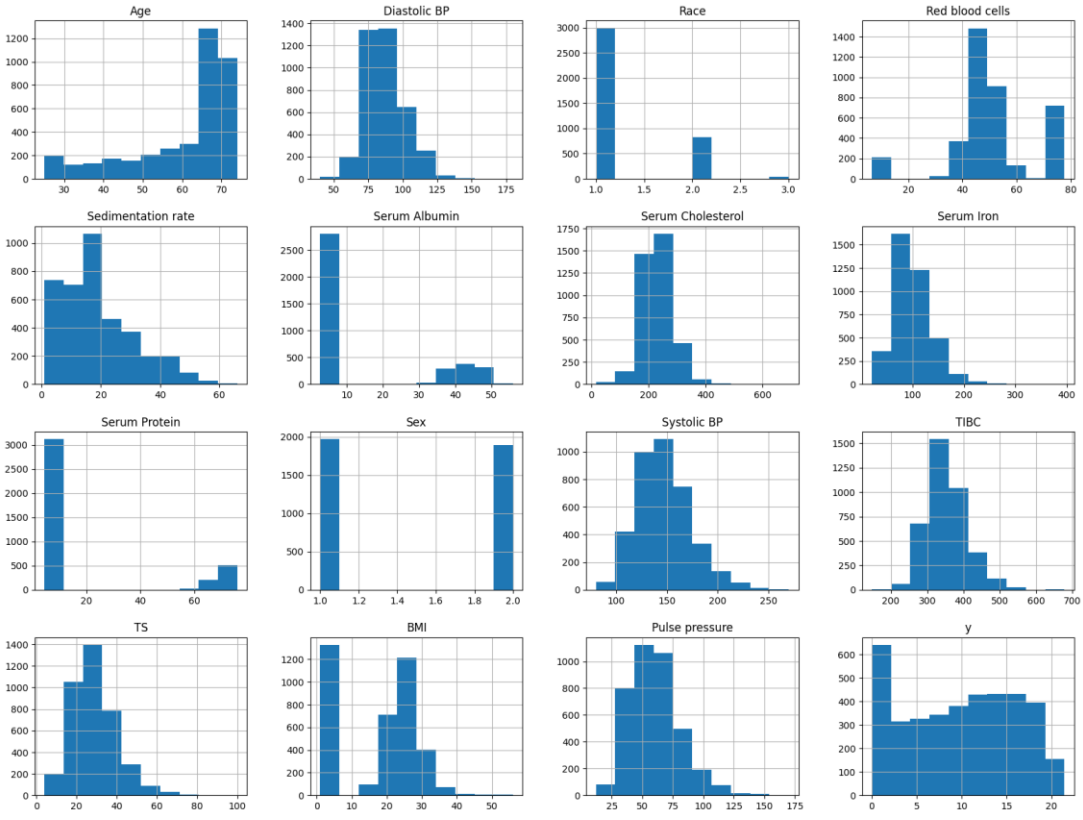


Gambar 4. 4 outlier

Hasil analisis outlier menunjukkan bahwa terdapat beberapa outlier pada beberapa variabel diantaranya yaitu *age*, *diastolic BP*, *Red Blood Cells*, *Sedimentation rate*, *serum cholesterol*, *serum iron*, *serum protein*, *systolic BP*, TIBC, TS, dan *pulse pressure*. Penanganan *outlier* tidak dilakukan karena penghapusan *outlier* akan mengurangi jumlah data yang tersisa secara signifikan, mengakibatkan hilangnya informasi yang berharga. Sementara itu, imputasi data tidak efektif dalam mendistribusikan data dengan baik dan masih menyisakan *outlier* pada beberapa variabel.

3) Distribusi data

Selanjutnya memeriksa distribusi data menggunakan metode visualisasi *histogram* untuk mendapatkan gambaran umum tentang distribusi nilai.



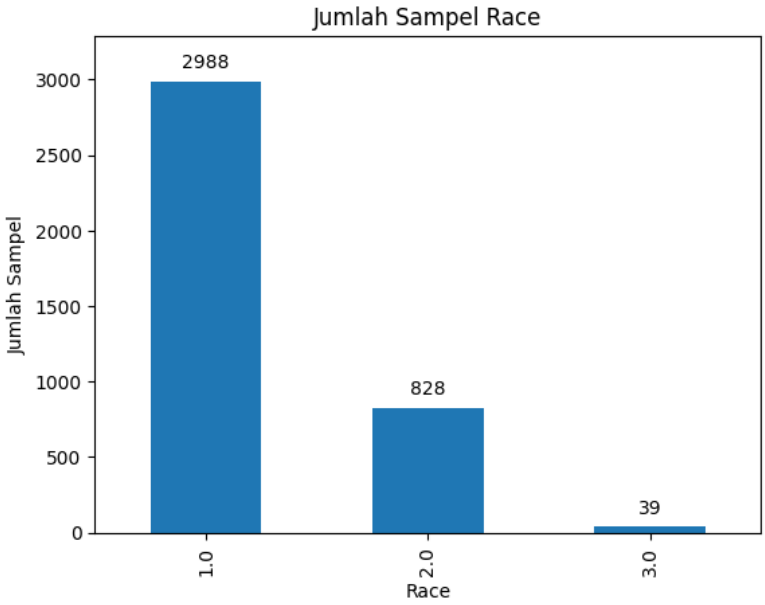
Gambar 4. 5 distribusi data histogram

Pada visualisasi distribusi diatas terlihat bahwa variabel-variabel tersebut tidak simetris atau tidak terdistribusi dengan baik. Sebagian besar variabel terdistribusi *Right-Skewed* yang artinya sebagian besar

nilai data terpusat di sisi kiri, sementara nilai-nilai ekstrem yang lebih besar menyebar ke arah kanan. Ini menunjukkan nilai rata-rata (*mean*) lebih besar daripada *median*. Dan juga ada beberapa variabel terdistribusi *Left-Skewed* yang berarti bahwa sebagian besar nilai data terpusat di sisi kanan, sementara nilai-nilai ekstrem yang lebih besar menyebar ke arah kiri. Ini menunjukkan nilai rata-rata (*mean*) lebih kecil daripada *median*.

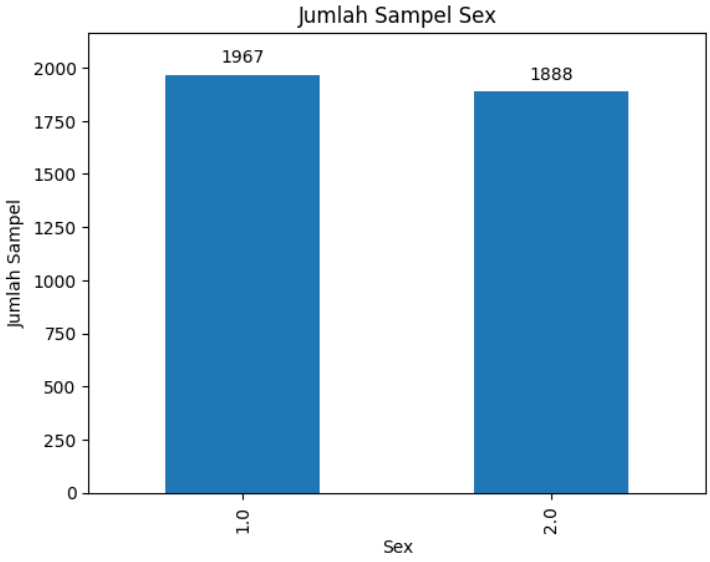
4) *Corelation analysis*

Untuk melihat hubungan antara masing-masing variabel perlu dilakukan *unvariate analysis*.



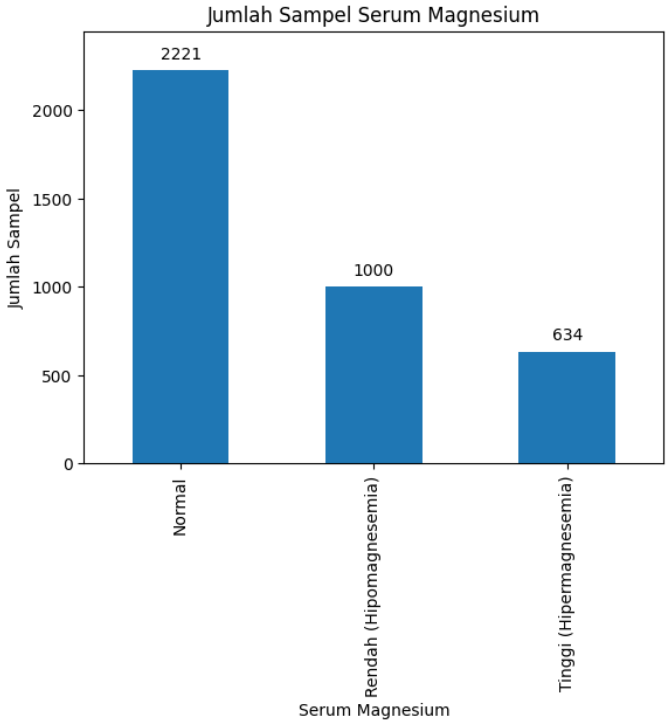
Gambar 4. 6 jumlah sample race

Variabel *race* menunjukkan bahwa distribusi jumlah sampel menunjukkan dominasi signifikan dari kelompok 1 yaitu kelompok dengan kulit berwarna putih, yang memiliki jumlah sampel terbanyak di antara semua kategori. Ini menunjukkan bahwa model lebih tergenalisir dengan *race* dari kelompok 1.



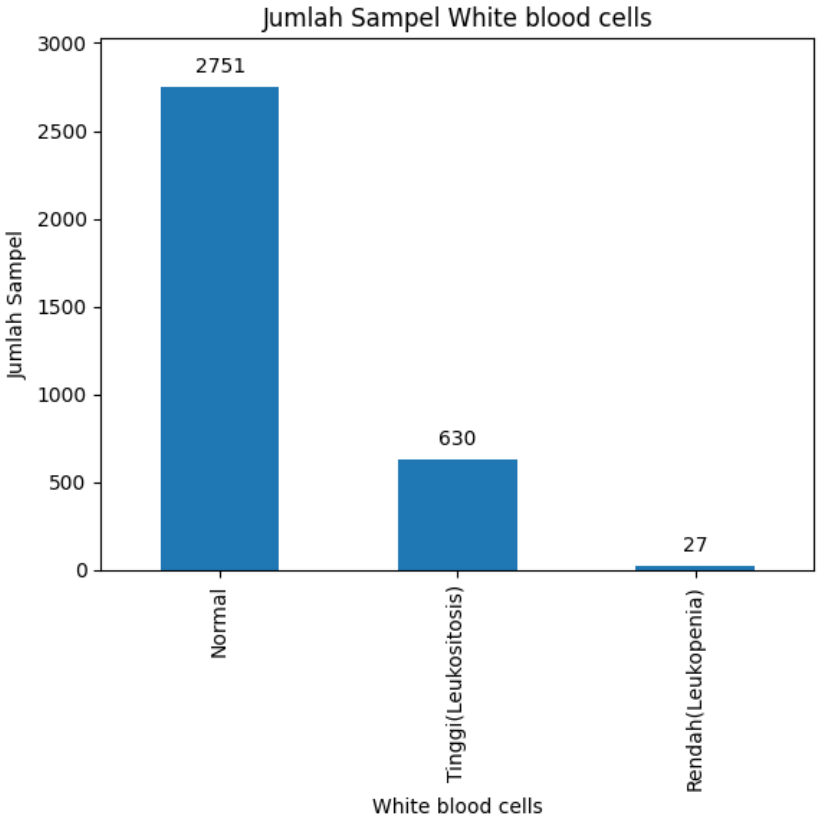
Gambar 4. 7 jumlah sample sex

Grafik ini memperlihatkan bahwa proporsi kelompok 1 yaitu pria yang secara signifikan lebih besar dibandingkan dengan kelompok 2 yaitu wanita. Ini menunjukkan bahwa penderita *cardiovascular* didominasi oleh pria, sehingga model lebih tergenalisir dengan kelompok *sex* pria.



Gambar 4. 8 jumlah sample serum magnesium

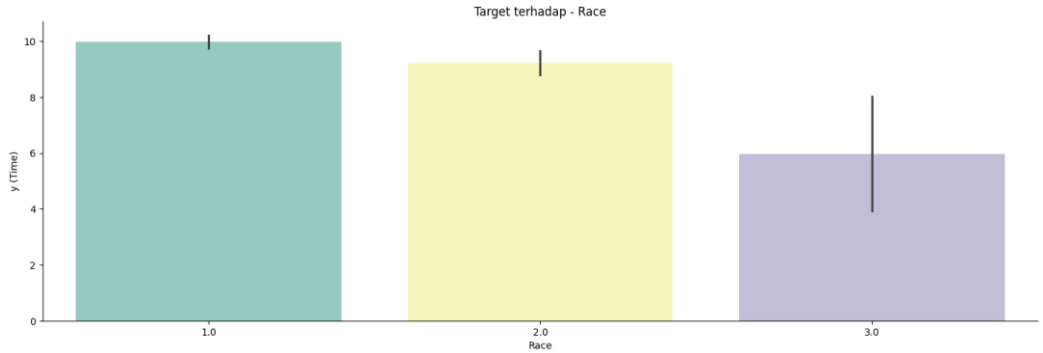
Ini menunjukkan bahwa jumlah sampel dengan kadar *serum magnesium* dalam kategori normal adalah yang terbanyak di antara semua kategori dan mengindikasikan fungsi tubuh yang optimal. Sehingga mengindikasikan bahwa sebagian besar individu memiliki kadar *serum magnesium* yang berada dalam rentang normal, sementara jumlah sampel dengan kadar yang rendah dan tinggi relatif lebih sedikit.



Gambar 4. 9 jumlah sample white blood cells

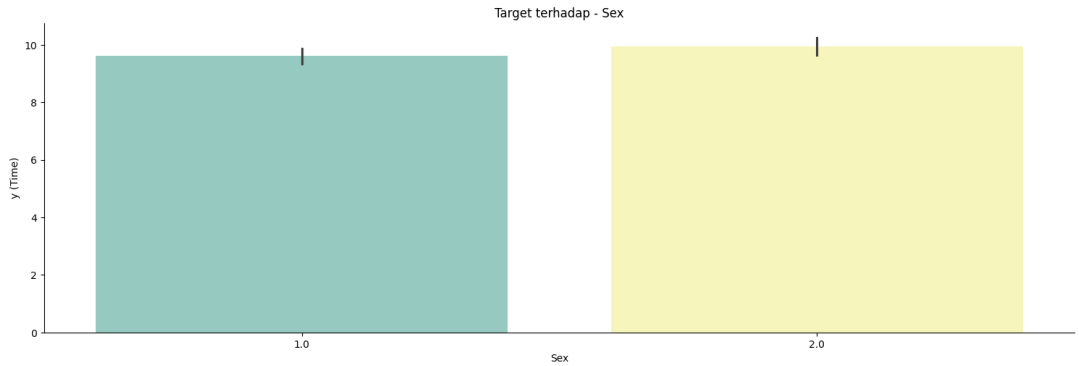
Grafik ini menunjukkan bahwa kadar WBC normal lebih umum, yang mungkin menunjukkan bahwa adanya gangguan atau kondisi medis yang mempengaruhi kadar WBC pada sebagian kecil populasi. Sementara itu, jumlah sampel dengan kadar WBC yang tinggi atau rendah dapat mengindikasikan adanya infeksi, peradangan, atau kondisi kesehatan lainnya adalah relatif sedikit. Kondisi ini mungkin menunjukkan bahwa dataset mencakup populasi yang mayoritasnya sehat atau tidak mengalami gangguan besar pada sistem kekebalan tubuh.

Adapun untuk melihat hubungan antara beberapa variabel dapat divisualisasikan sebagai berikut.



Gambar 4. 10 Race vs Y(time)

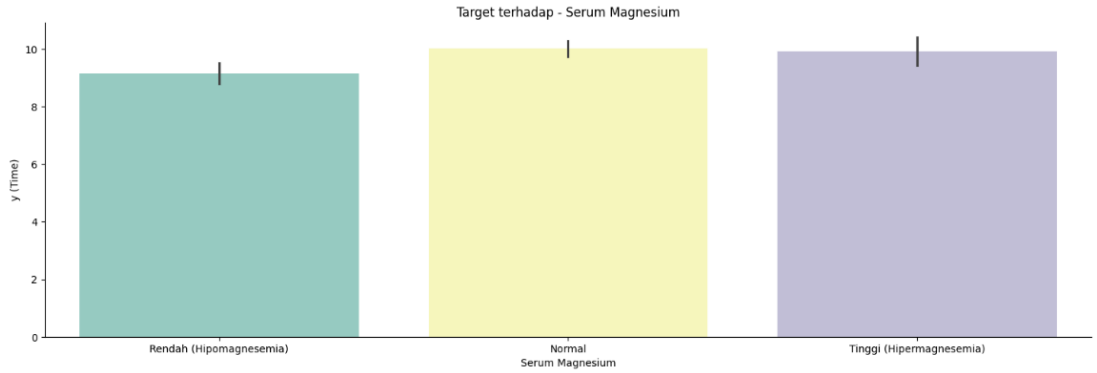
Grafik ini menunjukkan bahwa *race* dengan kelompok 1 atau kelompok kulit putih memiliki waktu kelangsungan hidup (variabel y atau waktu kelangsungan hidup pasien) yang bervariasi, dengan sejumlah individu menunjukkan waktu kelangsungan hidup yang lebih lama atau lebih pendek. Ini mengindikasikan bahwa *race* dengan kulit putih menunjukkan waktu kelangsungan hidup yang lebih panjang dibandingkan dengan kelompok lain. Kemungkinan ini dikarenakan perbedaan dalam akses ke perawatan kesehatan, kondisi sosial ekonomi, atau faktor genetik yang mempengaruhi hasil klinis.



Gambar 4. 11 Sex vs Y (time)

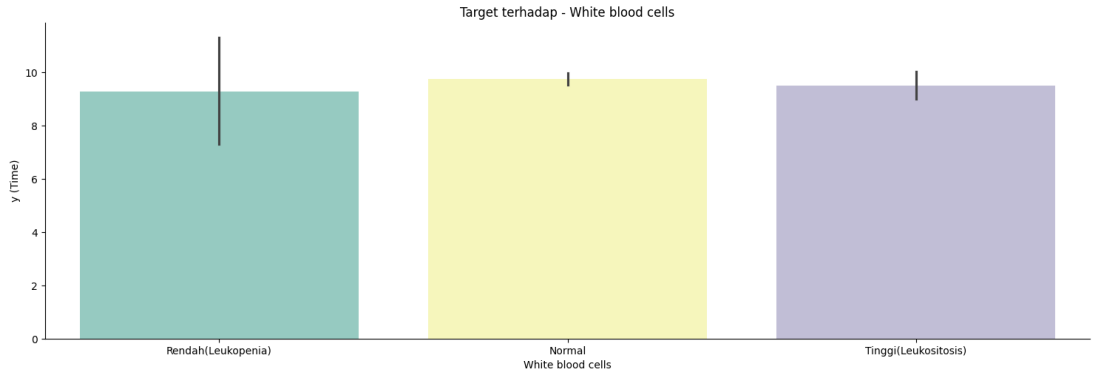
Data menunjukkan bahwa *sex* dengan nilai 2 atau wanita lebih mendominasi, namun perbedaan dalam waktu kelangsungan hidup antara wanita dan pria tidak terlalu signifikan. Ini berarti bahwa meskipun jenis kelamin wanita lebih mendominasi, faktor jenis kelamin

tidak memiliki dampak besar pada waktu kelangsungan hidup pasien secara keseluruhan.



Gambar 4. 12 Serum Magnesium vs Y (time)

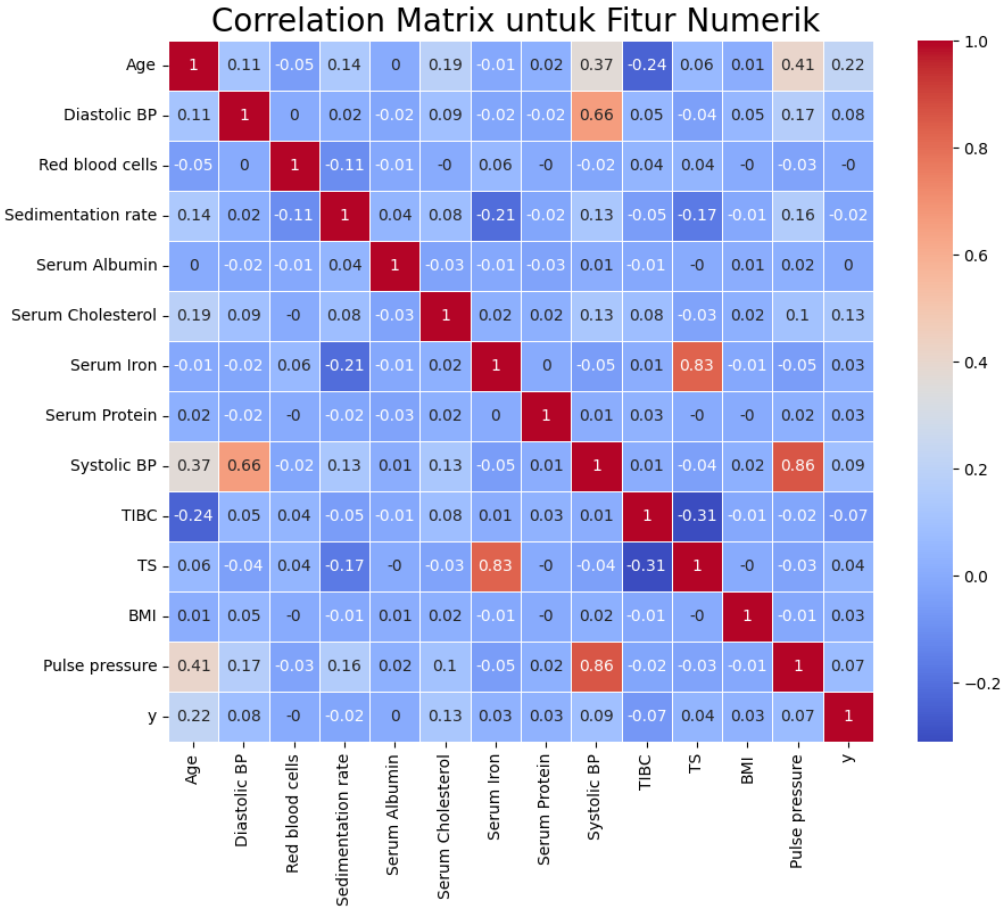
Sebagian besar pasien memiliki kadar *serum magnesium* yang berada dalam rentang normal. Namun kadar *magnesium* yang berada dalam batas normal tidak memiliki dampak yang signifikan terhadap waktu kelangsungan hidup dibandingkan dengan *kadar magnesium* yang lebih tinggi atau lebih rendah. Kadar *serum magnesium* yang normal tampaknya tidak berkorelasi secara signifikan dengan perubahan waktu kelangsungan hidup. Ini menunjukkan bahwa faktor-faktor lain mungkin lebih berpengaruh terhadap waktu kelangsungan hidup pasien dibandingkan dengan kadar *serum magnesium*.



Gambar 4. 13 White blood cells vs Y (time)

Kadar normal WBC mendominasi, namun perbedaan antara keitga kadar WBC tidak terlalu mencolok. Sebagian besar pasien yang memiliki kadar WBC yang normal, ini menunjukkan bahwa pasien

dengan kadar WBC dalam rentang normal cenderung memiliki waktu kelangsungan hidup yang relatif stabil. Hal ini mungkin menunjukkan bahwa kadar WBC normal dapat berhubungan dengan prognosis yang lebih baik atau stabilitas kesehatan. Lain halnya dengan pasien dengan kadar WBC yang sangat tinggi atau rendah dapat menunjukkan adanya kondisi medis seperti infeksi kronis, peradangan, atau gangguan hematologi, yang mungkin mempengaruhi waktu kelangsungan hidup mereka.



Gambar 4. 14 corelation matrix

Korelasi antara *systolic* BP dan *pulse pressure* adalah **0.86**, menunjukkan bahwa kedua variabel ini memiliki hubungan yang sangat kuat dan positif. Artinya, meningkatnya tekanan darah sistolik sering kali disertai dengan peningkatan dalam *pulse pressure*. ini

mengindikasikan bahwa perubahan dalam tekanan darah sistolik cenderung mempengaruhi *pulse pressure* secara signifikan.

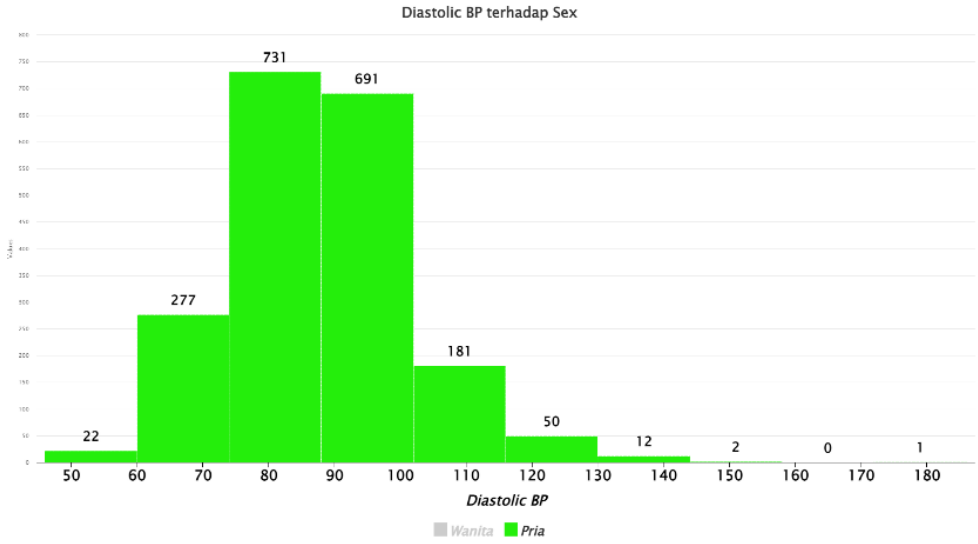
Korelasi antara *serum iron* dan total serum (TS) adalah **0.83**, menunjukkan hubungan kedua variabel ini kuat dan positif. Hal ini menunjukkan bahwa kadar *serum iron* dan TS cenderung bergerak seiring. Kadar *serum iron* yang lebih tinggi atau lebih rendah sering kali dipengaruhi oleh kadar TS.

Korelasi antara *systolic* BP dan diastolic blood pressure adalah **0.66**, yang menunjukkan hubungan positif yang moderat (lemah dan kuat). Ini berarti bahwa tekanan darah sistolik dan diastolik cenderung bergerak dalam arah yang sama, di mana perubahan pada satu komponen tekanan darah dapat mempengaruhi komponen lainnya.

Korelasi antara usia (*age*) dan *pulse pressure* adalah **0.41**, yang menunjukkan hubungan positif yang moderat. Hal ini dapat menunjukkan dengan seiring bertambahnya usia, ada kecenderungan dalam peningkatan *pulse pressure*, meskipun hubungan ini tidak terlalu kuat.

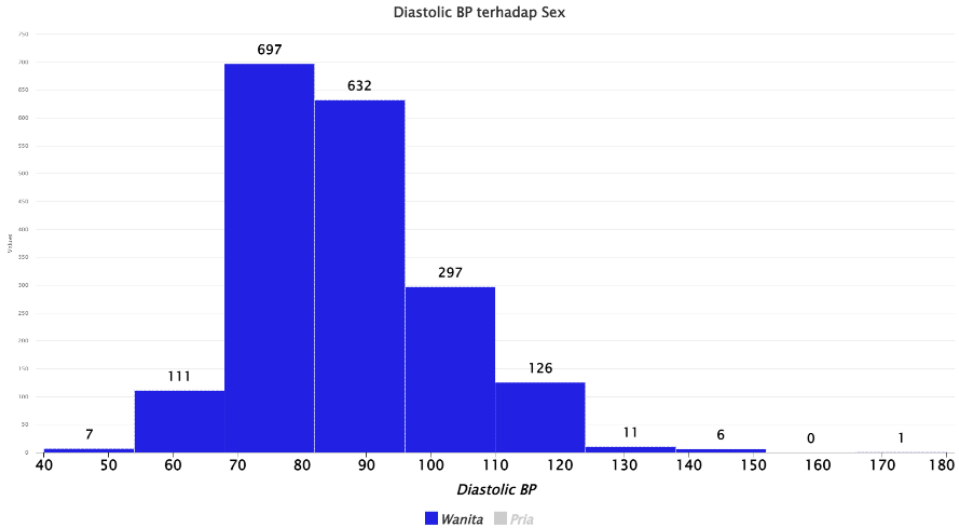
Korelasi antara usia (*age*) dan tekanan darah sistolik adalah **0.37**, ini menunjukkan hubungan positif yang lemah. Ini mengindikasikan bahwa meskipun usia dapat mempengaruhi tekanan darah sistolik.

Korelasi antara usia (*age*) dan variabel target (y) adalah **0.22**, ini menunjukkan hubungan positif yang lemah. Ini berarti bahwa hubungan antara usia dan variabel target (waktu kelangsungan hidup pasien) relatif tidak signifikan. Meskipun ada kecenderungan bahwa usia dapat mempengaruhi variabel target, kekuatan hubungan ini cukup rendah, kemungkinan faktor lain lebih berpengaruh terhadap variabel target.



Gambar 4. 15 Diastolic BP terhadap Sex Pria

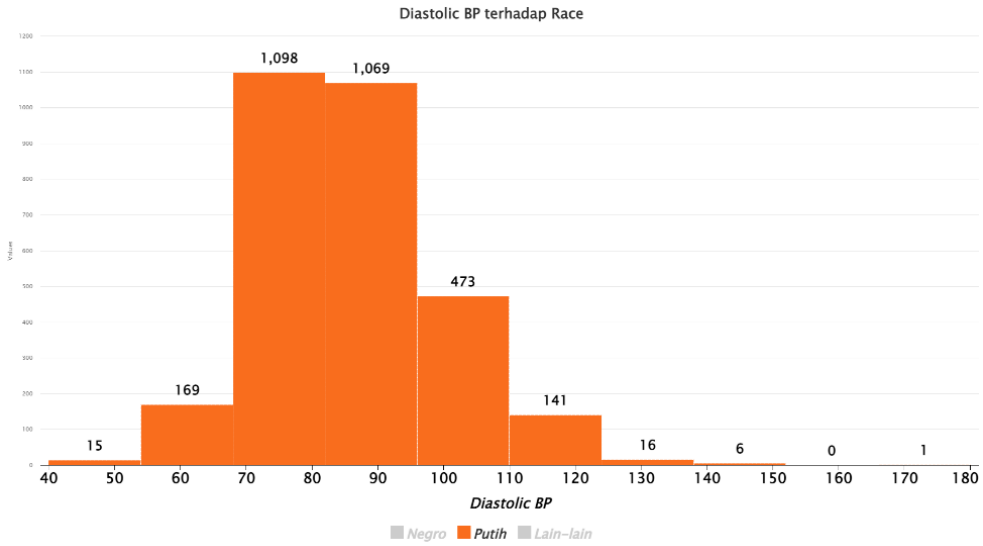
Gambar tersebut menunjukkan bahwa tekanan darah diastolik pada pria berkisar di sekitar angka 80 mmHg, ini dianggap tekanan darah diastolik yang normal. Meskipun masih dalam batas normal, angka ini mendekati ambang prehipertensi, yang bisa menjadi tanda peringatan untuk tindakan pencegahan lebih lanjut.



Gambar 4. 16 Diastolic BP terhadap Sex wanita

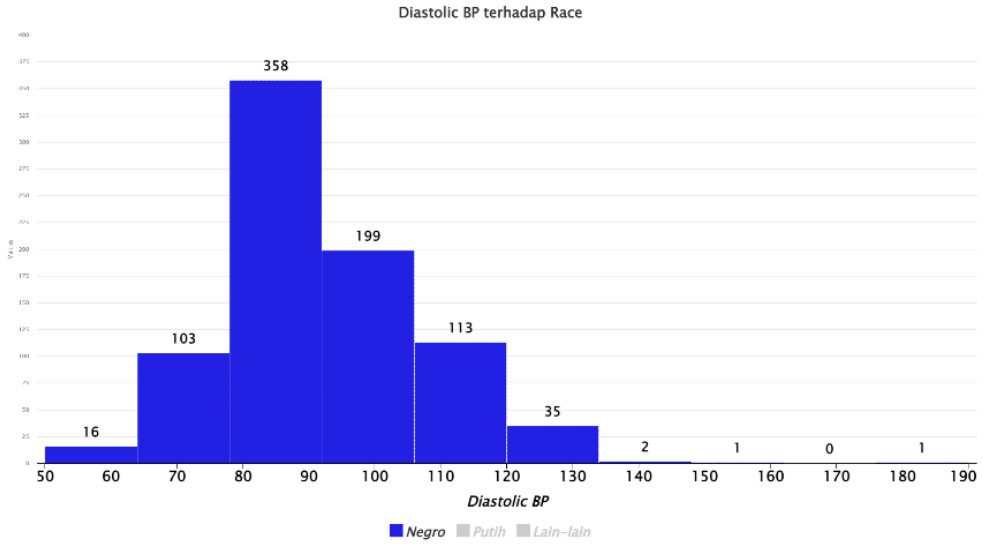
Gambar yang dihasilkan menunjukkan bahwa tekanan darah diastolik pada wanita paling banyak berada pada rentang 70 - 80 mmHg. Rentang ini untuk tekanan darah diastolik umumnya dianggap

dalam batas normal dan sehat. Sama halnya dengan tekanan diastolik pada pria, rentang ini mendekati ambang prehipertensi.



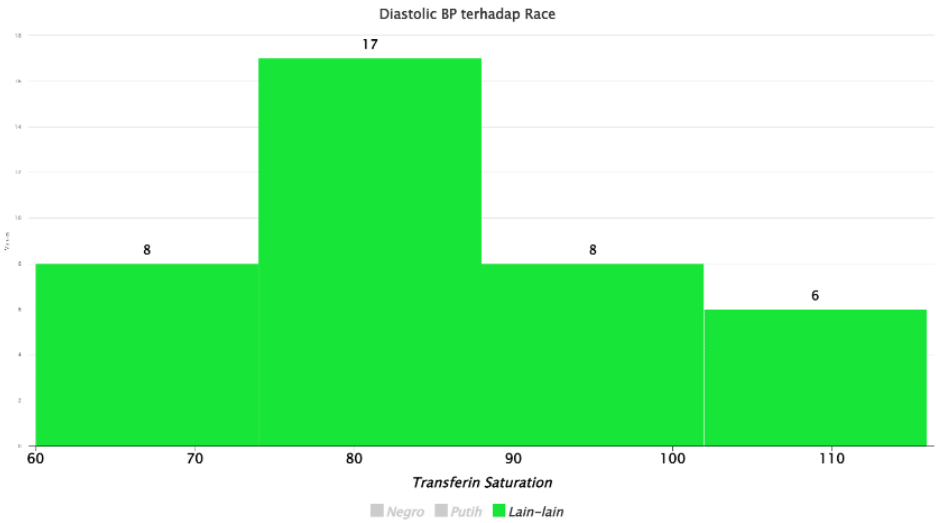
Gambar 4. 17 Diastolic BP terhadap Sex Race Putih

Pada gambar tersebut menunjukkan bahwa tekanan darah diastolik pada ras putih paling banyak berada pada rentang 70 - 80 mmHg. Rentang ini untuk tekanan darah diastolik umumnya dianggap dalam batas normal dan sehat. Namun tekanan darah diastolik yang berada > 80 juga cukup banyak yang menandakan tekanan darah ini cukup tinggi.



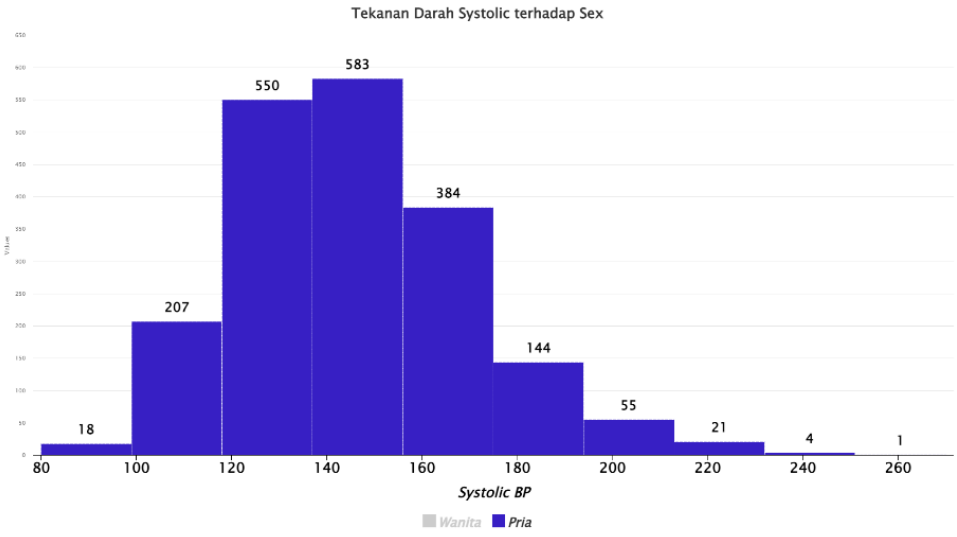
Gambar 4. 18 Diastolic BP terhadap Race Negro

Pada gambar tersebut menunjukkan bahwa tekanan darah diastolik pada ras negro paling banyak berada pada rentang 80 - 90 mmHg. Rentang ini menunjukkan ras ini memiliki tekanan darah diastolik yang tinggi dan mendekati hipertensi 1.



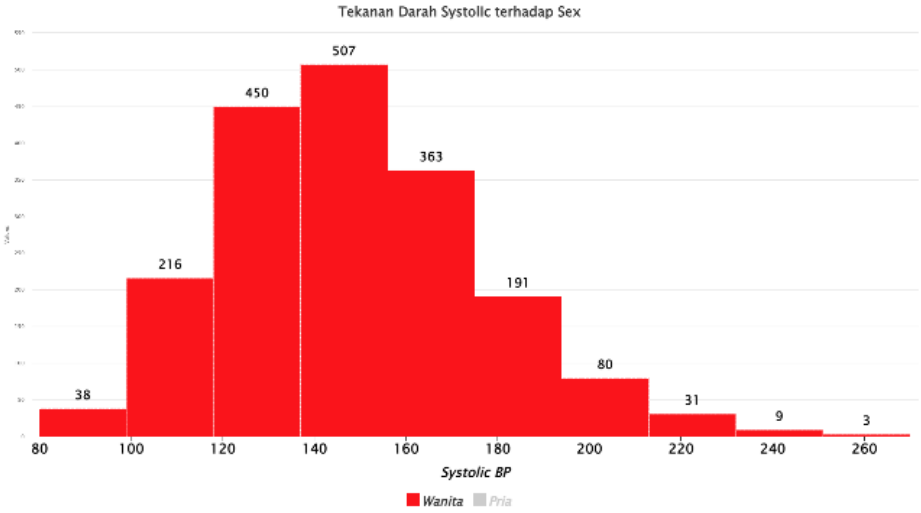
Gambar 4. 19 Diastolic BP terhadap Race lainnya

Gambar tersebut menunjukkan bahwa tekanan darah diastolik pada pada ras lainnya berkisar di sekitar angka 80 mmHg, ini dianggap tekanan darah diastolik yang normal. Namun angka ini mendekati ambang prehipertensi, yang bisa menjadi tanda peringatan untuk tindakan pencegahan lebih lanjut.



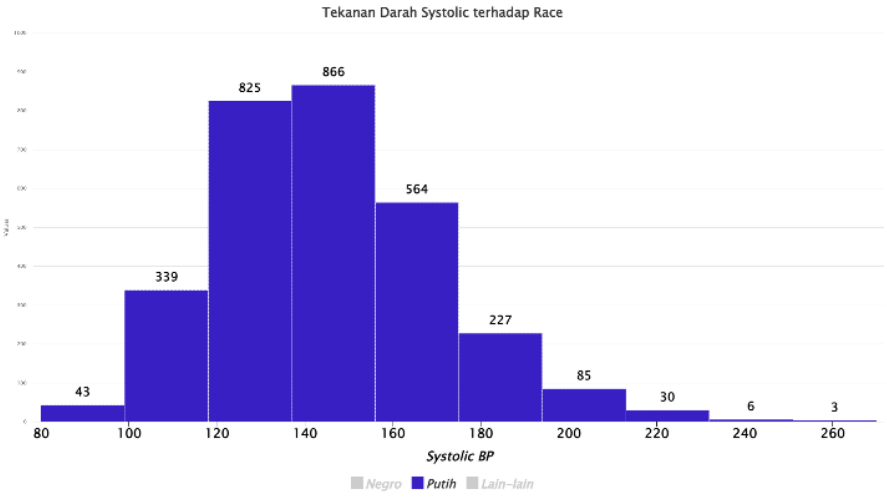
Gambar 4. 20 Systolic BP terhadap Sex Pria

Pada gambar tersebut menunjukkan bahwa Tekanan Darah *Systolic* pada pria terbanyak berada pada rentang 140 – 160 mmHg dimana ini menunjukkan prevalensi hipertensi di antara populasi pria, dan diklasifikasikan sebagai hipertensi tahap 2. Oleh karena itu, pria dalam rentang ini berisiko lebih tinggi mengalami komplikasi terkait hipertensi seperti penyakit jantung, *stroke*, dan kerusakan ginjal.



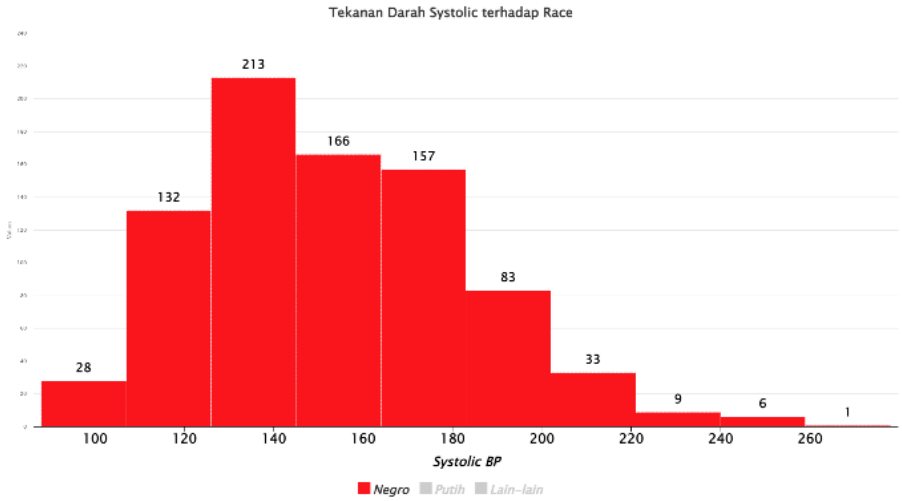
Gambar 4. 21 Systolic BP terhadap Sex Wanita

Pada gambar tersebut menunjukkan bahwa tekanan darah *Systolic* pada Wanita terbanyak berada pada rentang 140 – 160 mmHg yang sama dengan rentang tekanan darah *Systolic* pada Pria, dan diklasifikasikan sebagai hipertensi tahap 2.



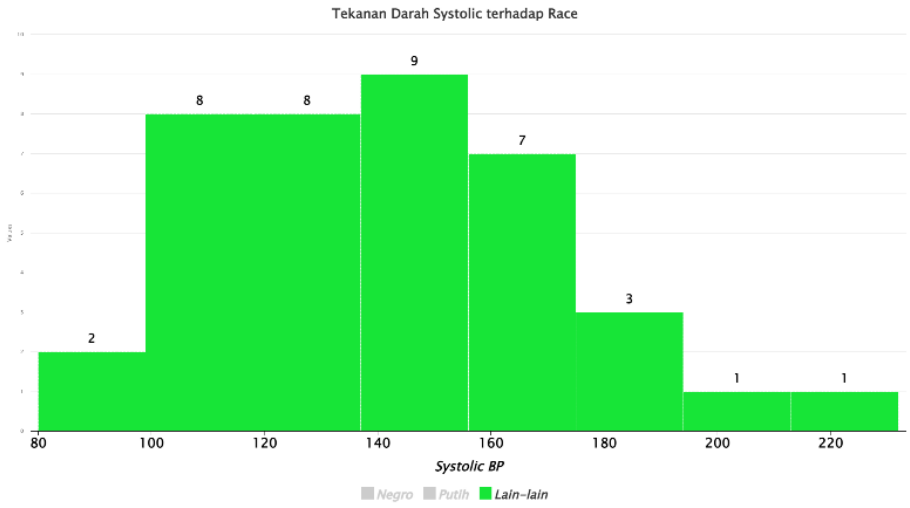
Gambar 4. 22 Systolic BP terhadap Rcae Putih

Ini menunjukkan bahwa Tekanan Darah *Systolic* pada ras putih terbanyak berada pada rentang 140 – 160 mmHg dimana ini menunjukkan bahwa banyak individu dari ras putih mungkin berada pada risiko tinggi untuk kondisi terkait hipertensi, seperti penyakit jantung, stroke, dan penyakit ginjal.



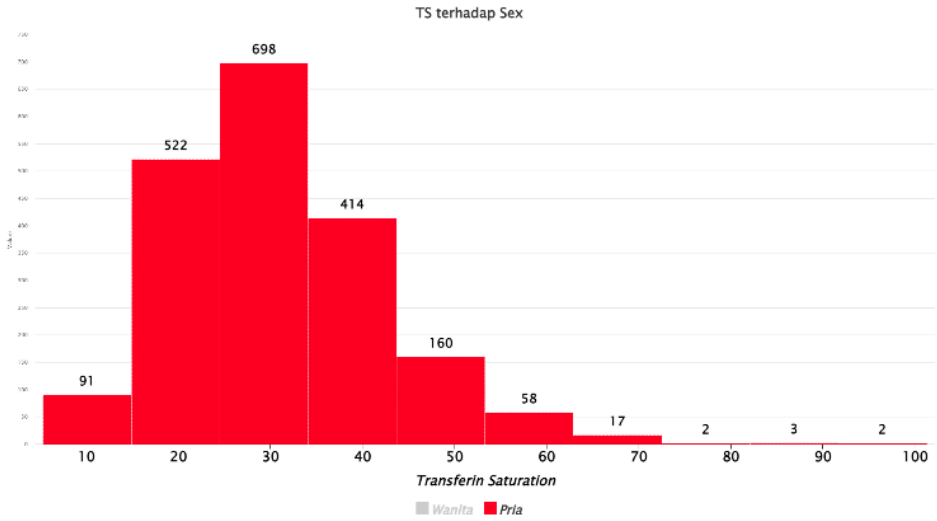
Gambar 4. 23 Systolic BP terhadap Race Negro

Ini menunjukkan bahwa Tekanan Darah *Systolic* pada ras putih terbanyak berada pada rentang 140 mmHg, dimana ini menunjukkan banyak individu dari ras negro memiliki Tekanan Darah *Systolic* yang relatif tinggi dan memiliki risiko lebih tinggi untuk kondisi seperti penyakit jantung, stroke, dan masalah *cardiovascular* lainnya.



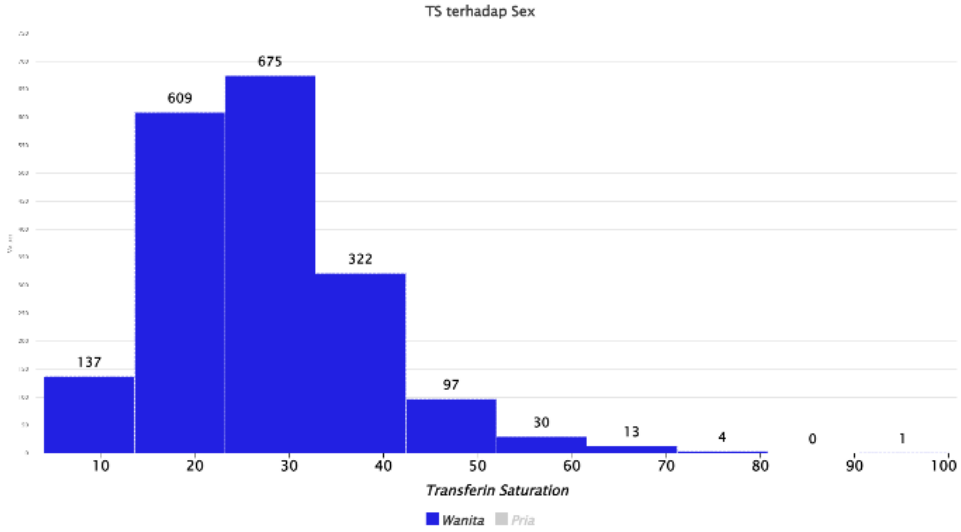
Gambar 4. 24 Systolic BP terhadap Race lainnya

Ini menunjukkan bahwa sebagian besar individu dari kelompok ras lainnya memiliki Tekanan Darah *Systolic* terbanyak berada pada rentang 140 - 160 mmHg, dimana ini menunjukkan banyak individu dalam kelompok ini berisiko lebih tinggi terhadap kondisi terkait hipertensi.



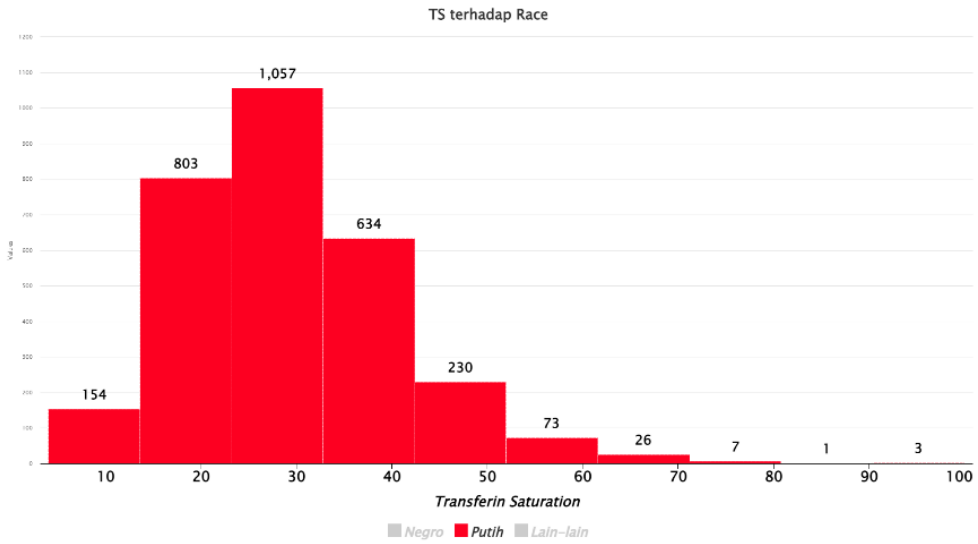
Gambar 4. 25 TS terhadap Sex Pria

Gambar tersebut menunjukkan bahwa nilai saturasi transferin pada pria paling banyak berada pada rentang sekitar 30. Hal ini mengindikasikan bahwa mayoritas pria dalam dataset memiliki tingkat saturasi transferin yang mendekati atau di sekitar angka 30.



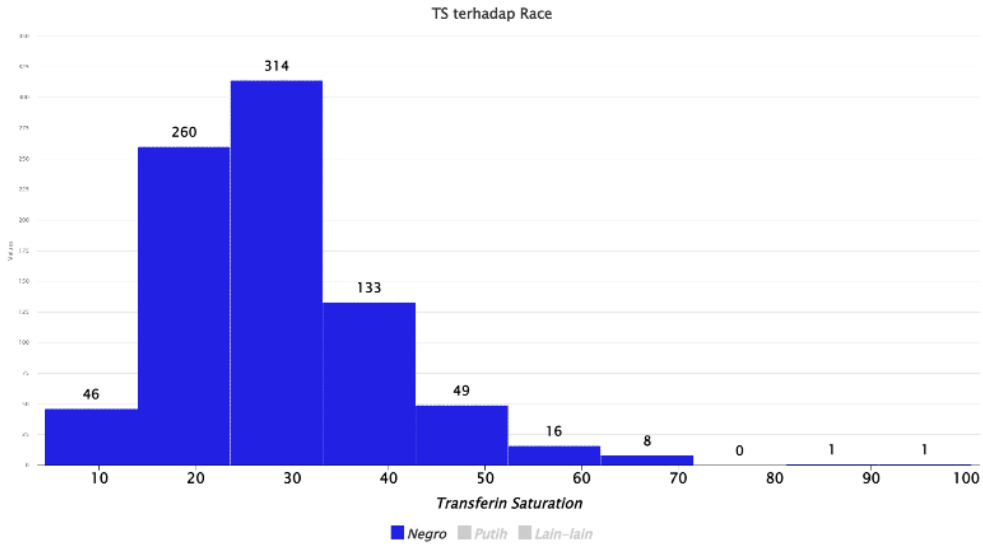
Gambar 4. 26 TS terhadap Sex Wanita

Gambar tersebut menunjukkan bahwa nilai saturasi transferin pada wanita paling banyak berada pada rentang sekitar 30, ini menunjukkan hal yang sama dengan saturasi transferrin pada pria.



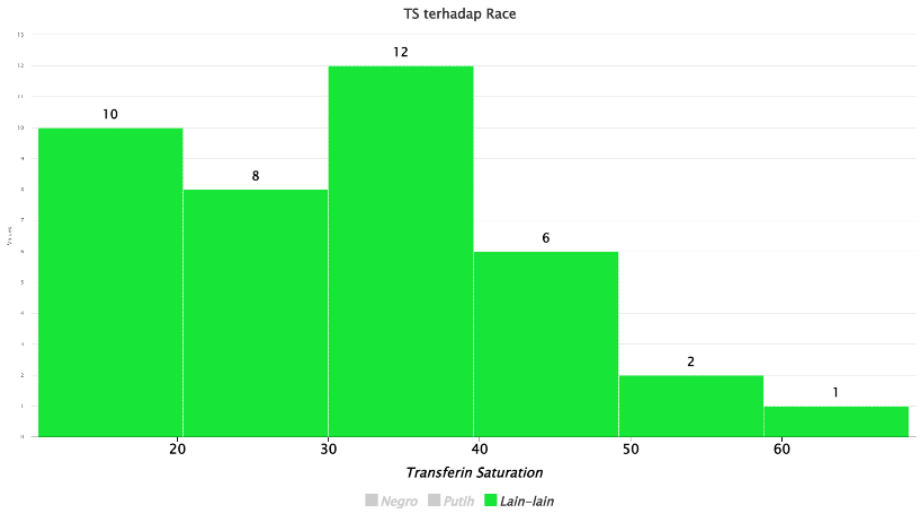
Gambar 4. 27 TS terhadap Race Putih

Ini menunjukkan bahwa nilai saturasi transferin pada ras putih paling banyak berada pada rentang sekitar 30. Hal ini mengindikasikan bahwa mayoritas ras putih dalam dataset memiliki tingkat saturasi transferin yang mendekati atau di sekitar angka 30.



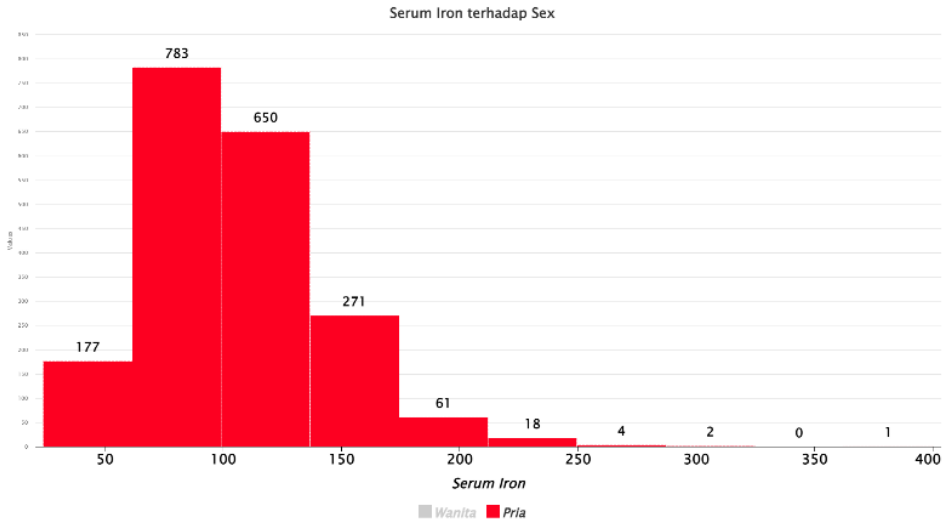
Gambar 4. 28 TS terhadap Race Negro

Gambar tersebut menunjukkan bahwa nilai saturasi transferin pada ras negro paling banyak berada pada rentang sekitar 30, ini menunjukkan hal yang sama dengan saturasi transferrin pada ras putih.



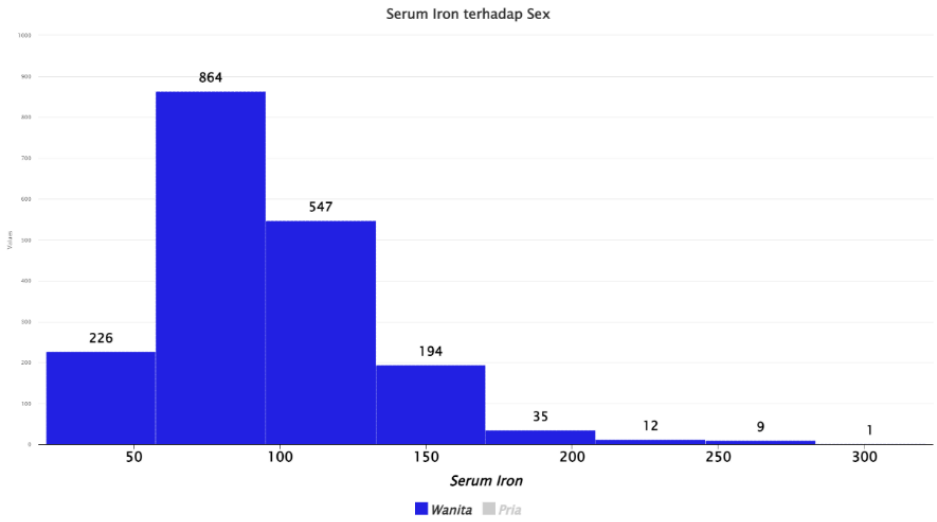
Gambar 4. 29 TS terhadap Race lainnya

Ini menunjukkan bahwa sebagian besar individu dari kelompok ras lainnya memiliki saturasi transferin pada rentang 30 – 40.



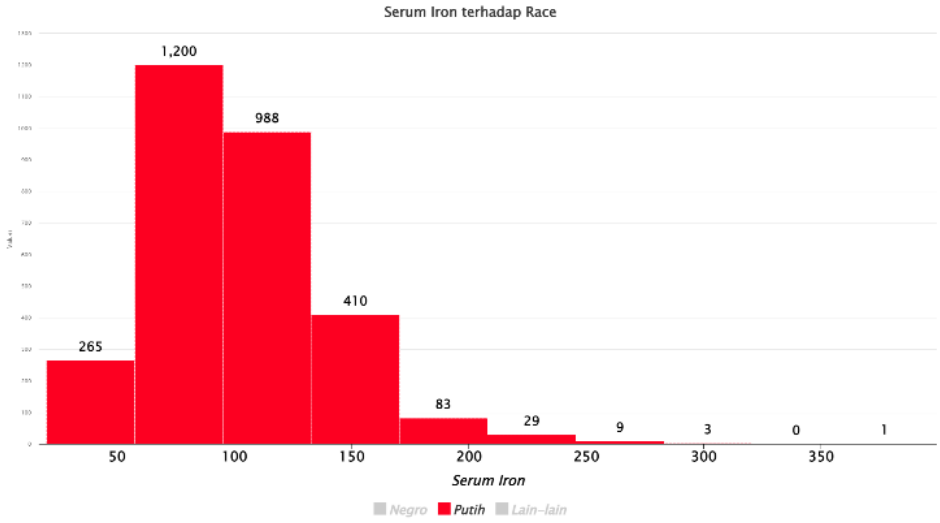
Gambar 4. 30 Serum Iron terhadap Sex Pria

Ini menunjukkan bahwa nilai serum *iron* pada pria cenderung lebih tinggi dan terkonsentrasi di sekitar rentang 50 – 100 mmHg. Ini menunjukkan bahwa mayoritas pria dalam dataset memiliki level serum *iron* yang cukup tinggi.



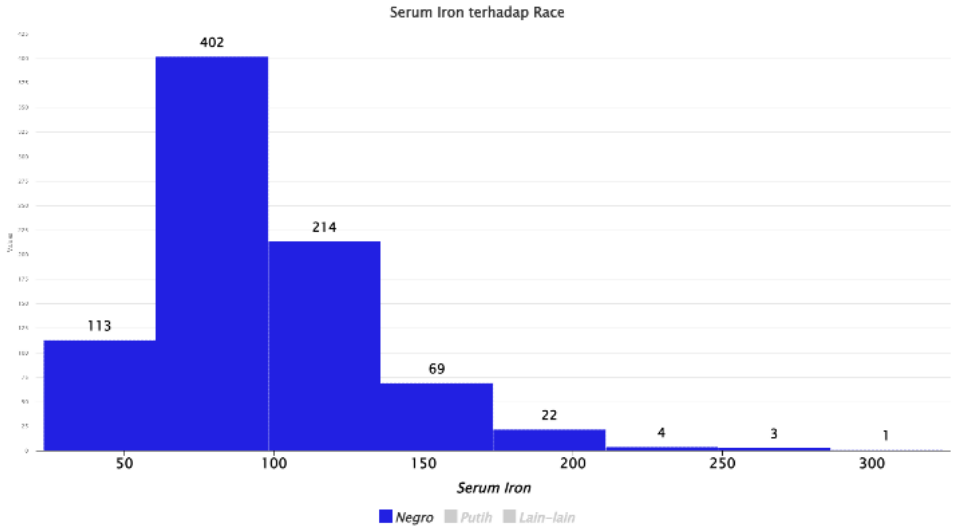
Gambar 4. 31 Serum Iron terhadap Sex Wanita

Ini menunjukkan bahwa nilai serum *iron* pada Wanita juga berada di sekitar rentang 50 – 100 mmHg. Ini menunjukkan bahwa mayoritas wanita memiliki level serum *iron* yang cukup tinggi.



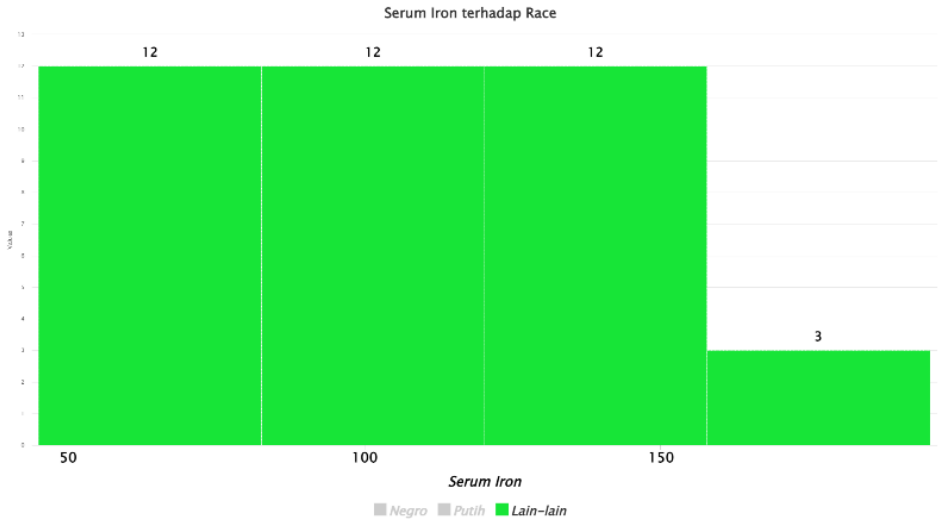
Gambar 4. 32 Serum Iron terhadap Race Putih

Ini menunjukkan bahwa nilai serum *iron* pada ras putih paling banyak berada pada rentang sekitar 50 - 100 mmHg. Hal ini mengindikasikan bahwa mayoritas ras putih dalam dataset memiliki tingkat serum iron pada rentang tersebut.



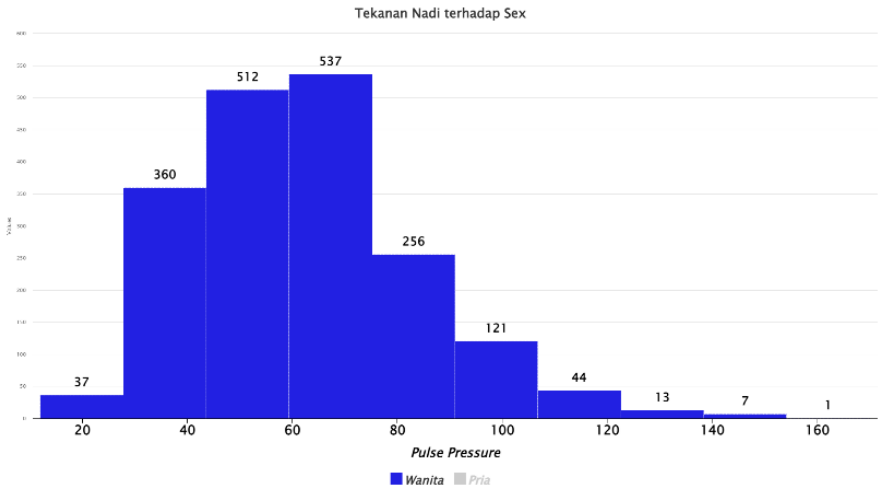
Gambar 4. 33 Serum Iron terhadap Race Negro

Gambar tersebut menunjukkan bahwa nilai serum *iron* pada ras negro paling banyak berada pada rentang sekitar 50 – 100 mmHg, ini menunjukkan hal yang sama dengan serum *iron* pada ras putih.



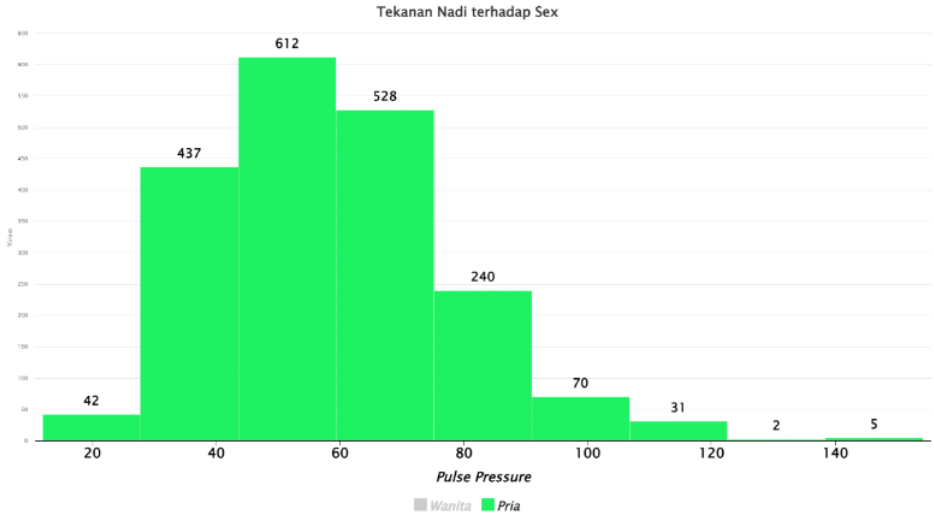
Gambar 4. 34 Serum Iron terhadap Race lainnya

Distribusi serum iron pada ras ini terlihat merata di antara ketiga rentang tersebut. Ini menunjukkan bahwa variasi serum iron di antara individu-individu dari ras lainnya cukup terbatas.



Gambar 4. 35 Tekanan Nadi terhadap Sex Wanita

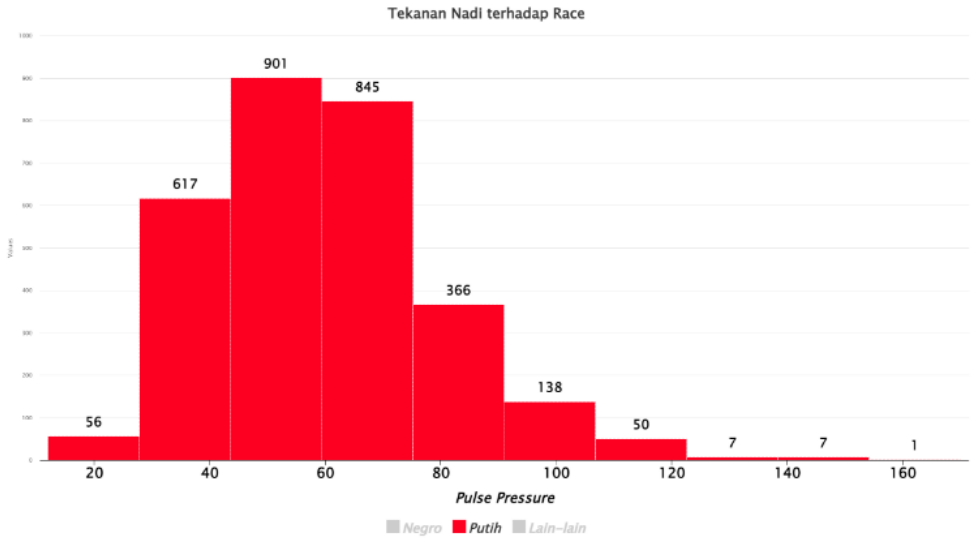
Pada gambar tersebut terlihat bahwa tekanan nadi (*pulse pressure*) pada wanita cenderung berada pada rentang 60 - 80 mmHg. Ini mengindikasikan bahwa sebagian besar wanita dalam dataset ini memiliki tekanan nadi yang tinggi dalam kisaran tersebut, dan berisiko risiko terhadap kondisi seperti hipertensi, penyakit jantung koroner, dan gagal jantung.



Gambar 4. 36 Tekanan Nadi terhadap Sex pria

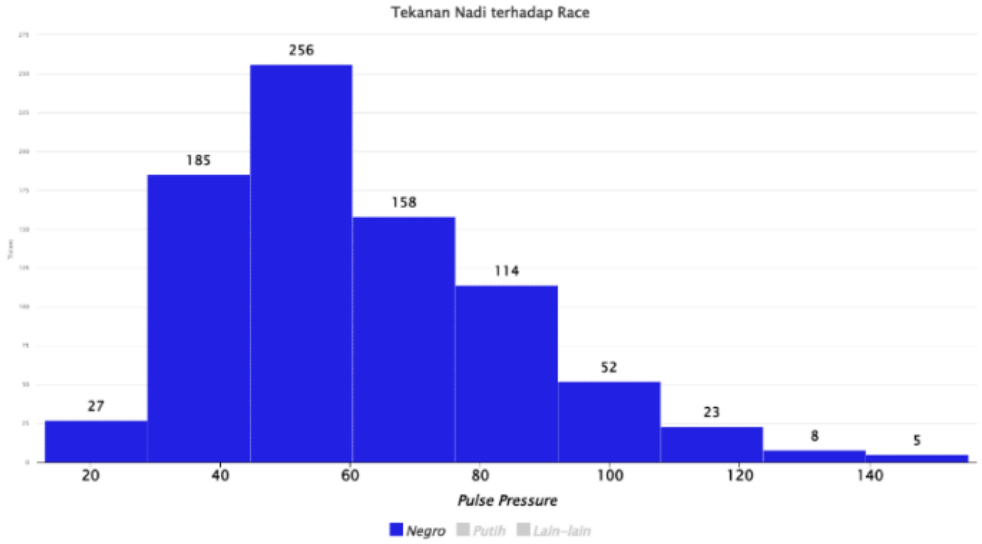
Pada gambar tersebut terlihat bahwa tekanan nadi (*pulse pressure*) pada pria cenderung berada pada rentang 40 - 60 mmHg. Ini mengindikasikan bahwa sebagian besar pria dalam dataset ini memiliki tekanan nadi yang normal, yang menunjukkan bahwa ada

keseimbangan yang baik antara tekanan darah sistolik dan diastolik, yang dapat mengindikasikan jantung yang sehat.



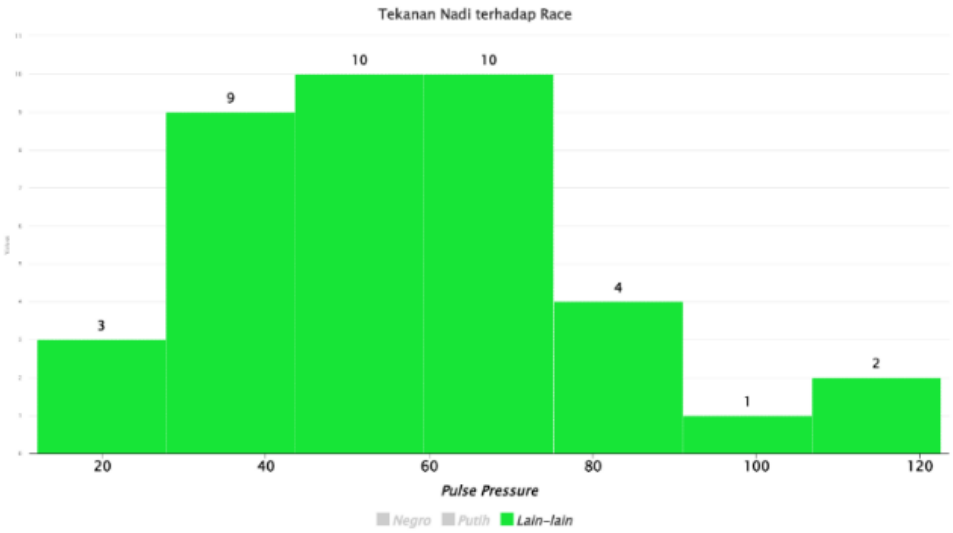
Gambar 4. 37 Tekanan Nadi terhadap Race Putih

Pada gambar tersebut terlihat bahwa tekanan nadi (*pulse pressure*) pada ras putih cenderung normal karena berada pada rentang 40 - 60 mmHg.



Gambar 4. 38 Tekanan Nadi terhadap Race Negro

Sama halnya dengan ras putih, ras ini memiliki tekanan nadi yang normal karena berada pada rentang 40 – 60 mmHg.



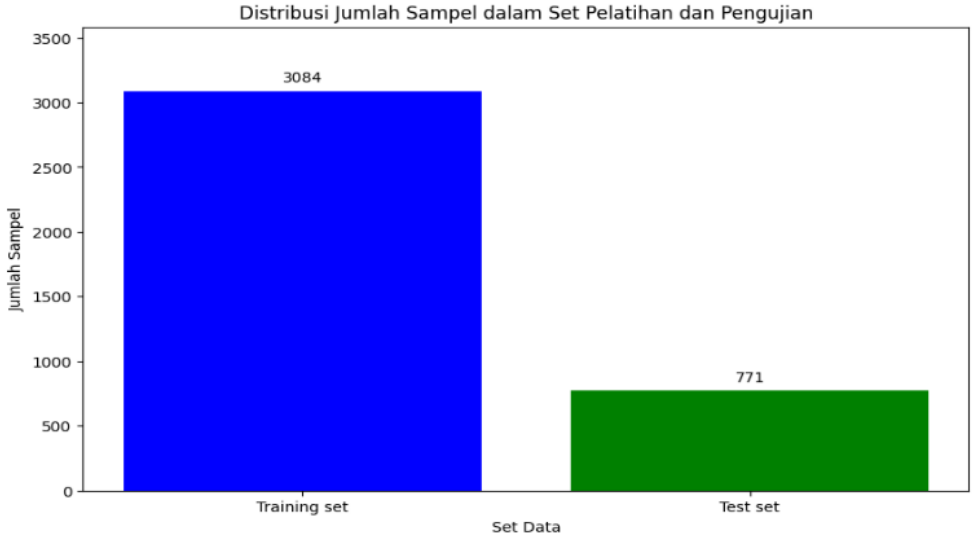
Gambar 4. 39 Tekanan Nadi terhadap Race lainnya

Dalam gambar tersebut, tekanan nadi yang sama antara ras lainnya menunjukkan bahwa ada keseragaman dalam rentang tekanan nadi di berbagai kelompok ras. Tekanan nadi pada ras lainnya yang sama berada dalam rentang yang bervariasi yaitu rendah, normal dan tinggi.

c. Data preprocessing

1) Splitting data

Dataset akan dibagi untuk *training data* dan *test data* dengan rasio sebesar 80:20. Pembagian *subset* ini dilakukan untuk melatih model dan menguji model.



Gambar 4. 40 splitting data

2) *Encoding fitur*

Pada tahapan ini dilakukan proses untuk mengubah setiap fitur kategori menjadi kolom biner baru dengan nilai 1 atau 0.

	Age	Diastolic BP	Race	Red blood cells	Sedimentation rate	Serum Albumin	Serum Cholesterol	Serum Iron	Serum Magnesium	Serum Protein	Sex	Systolic BP	TIBC	TS	White blood cells	BMI	Pulse pressure
0	65.0	98.0	2	53.0	3.0	4.4	338.0	130.0	0	66.0	0	158.0	364.0	35.7	0	25.0	60.0
1	72.0	96.0	2	77.7	48.0	4.2	137.0	78.0	0	6.7	1	170.0	325.0	24.0	0	2.0	74.0
2	70.0	90.0	2	45.5	29.0	4.2	227.0	67.0	1	66.0	1	160.0	355.0	18.9	0	27.0	70.0
3	74.0	80.0	2	56.1	4.0	4.0	301.0	57.0	1	6.8	0	130.0	344.0	16.6	2	2.0	50.0
4	53.0	70.0	2	77.7	27.0	4.2	276.0	65.0	0	6.7	1	142.0	342.0	19.0	0	18.0	72.0

Gambar 4. 41 encoding fitur category

Fitur-fitur *category* seperti *race*, *serum magnesium*, *sex* dan *white blood cells* telah diubah menjadi kolom biner. Masing-masing kolom ini akan berisi nilai 0, 1 ataupun 2 menyesuaikan dengan jumlah kategori masing-masing fitur tersebut.

3) Normalisasi data

Selanjutnya akan dilakukan penskalakan fitur atau variabel sehingga berada dalam rentang yang sama atau memiliki distribusi yang seragam, agar mempermudah proses pelatihan model.

	Age	Diastolic BP	Red blood cells	Sedimentation rate	Serum Albumin	Serum Cholesterol	Serum Iron	Serum Protein	Systolic BP	TIBC	TS	BMI	Pulse pressure
count	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000	3084.0000
mean	0.7212	0.3213	0.6228	0.2731	0.2237	0.3183	0.2131	0.2032	0.3488	0.3636	0.2588	0.2938	0.3034
std	0.2758	0.1028	0.2266	0.1893	0.3160	0.0812	0.0997	0.3537	0.1436	0.1107	0.1175	0.2094	0.1326
min	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
25%	0.5918	0.2609	0.5288	0.1385	0.0300	0.2670	0.1463	0.0268	0.2526	0.2877	0.1802	0.0364	0.2025
50%	0.8367	0.3116	0.5891	0.2308	0.0319	0.3129	0.1995	0.0324	0.3263	0.3523	0.2417	0.3636	0.2911
75%	0.9184	0.3841	0.6746	0.3846	0.6623	0.3651	0.2666	0.0423	0.4316	0.4247	0.3208	0.4545	0.3687
max	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Gambar 4. 42 normalisasi data

Semua variabel memiliki jumlah observasi yang sama, yaitu 3084. Ini menunjukkan bahwa tidak ada nilai yang hilang (*missing values*) dalam variabel-variabel ini setelah *preprocessing*. Semua variabel memiliki nilai minimum 0 dan maksimum 1. Ini menunjukkan bahwa data telah dinormalisasi atau diubah menjadi skala antara 0 dan 1.

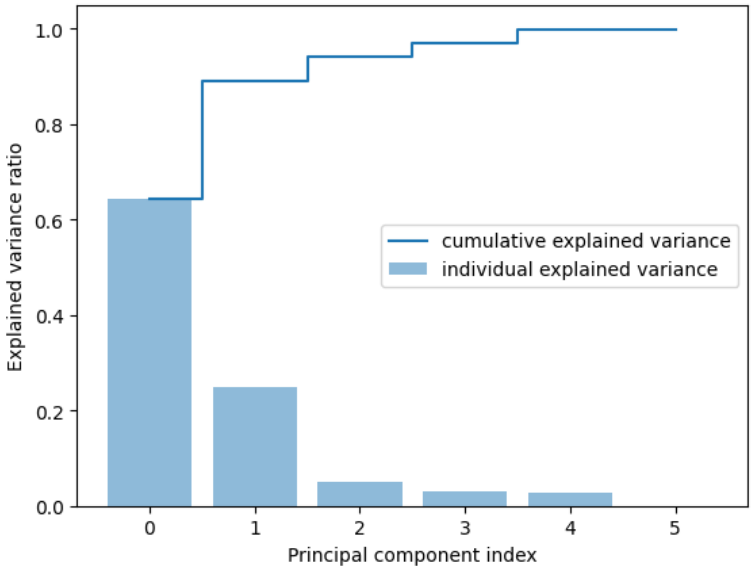
4) *Dimensional reduction*

Proses selanjutnya dilakukan pengurangan jumlah fitur (dimensi) dalam dataset tanpa menghilangkan informasi yang penting. Ini dilakukan jika data memiliki banyak fitur, yang sering kali menyebabkan masalah *overfitting*. Terdapat 4 kelompok *variable* berdasarkan kekuatan relasi masing-masing *variable*.

Tabel 4. 2 PCA

Nama PCA	Fitur
Pc_1	1. Age 2. Diastolic BP 3. Systolic BP 4. Pulse Pressure 5. Serum Cholestrol 6. BMI
Pc_2	1. Red Blood Cells 2. Sedimentation Rate
Pc_3	1. Serum Albumin 2. Serum Protein
Pc_4	1. Serum Iron 2. TS 3. TIBC

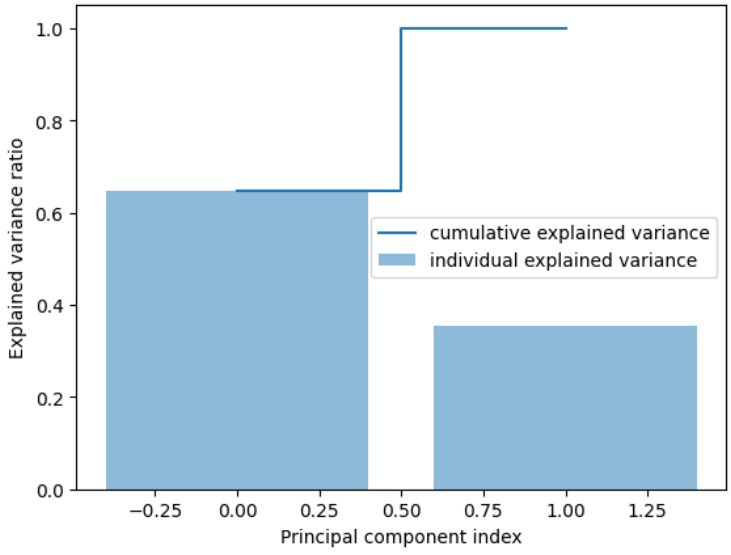
Variabel-variabel tersebut dipilih dan dibentuk berdasarkan analisis korelasi, dengan hubungan terkuat satu sama lain. Sehingga memastikan bahwa model yang dibangun dapat menangkap interaksi yang signifikan antara fitur-fitur tersebut dan meningkatkan akurasi serta pemahaman terhadap pola data yang ada.



Gambar 4. 43 pca_1

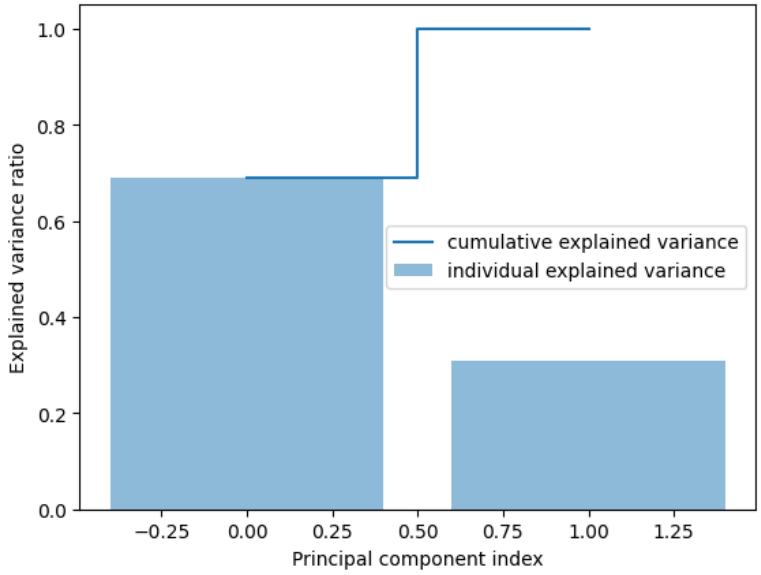
Pada gambar diatas dilakukan analisis komponen utama (PCA) pada *subset* data yang telah dipilih, dengan menghitung *individual explained variance* atau proporsi varians yang dijelaskan oleh setiap komponen utamanya.

Dan juga menghitung *cumulative explained variance* yaitu menghitung seberapa banyak varians dalam data yang dapat dijelaskan secara kumulatif oleh beberapa komponen utama. Dimana terdapat 6 variabel diagregasi menjadi 1 fitur baru bernama pca_1.



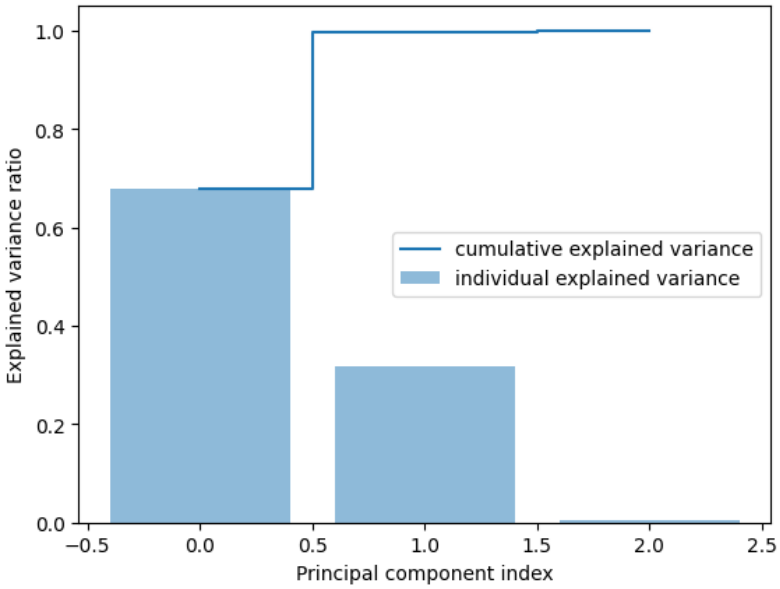
Gambar 4. 44 pca_2

Pada gambar diatas menunjukkan terdapat 2 variabel yang diagregasi menjadi 1 fitur baru bernama pca_2.



Gambar 4. 45 pca_3

Gambar tersebut menunjukkan terdapat 2 variabel yang diagregasi menjadi 1 fitur baru bernama pca_3.



Gambar 4. 46 pca_4

Pada Gambar tersebut didapatkan 3 variabel yang diagregasi menjadi 1 fitur baru bernama pca_4.

Setelah menentukan jumlah komponen PCA yang akan digunakan, lalu dilakukan transformasi *subset* data ke dalam komponen utama dengan jumlah komponen yang telah ditentukan.

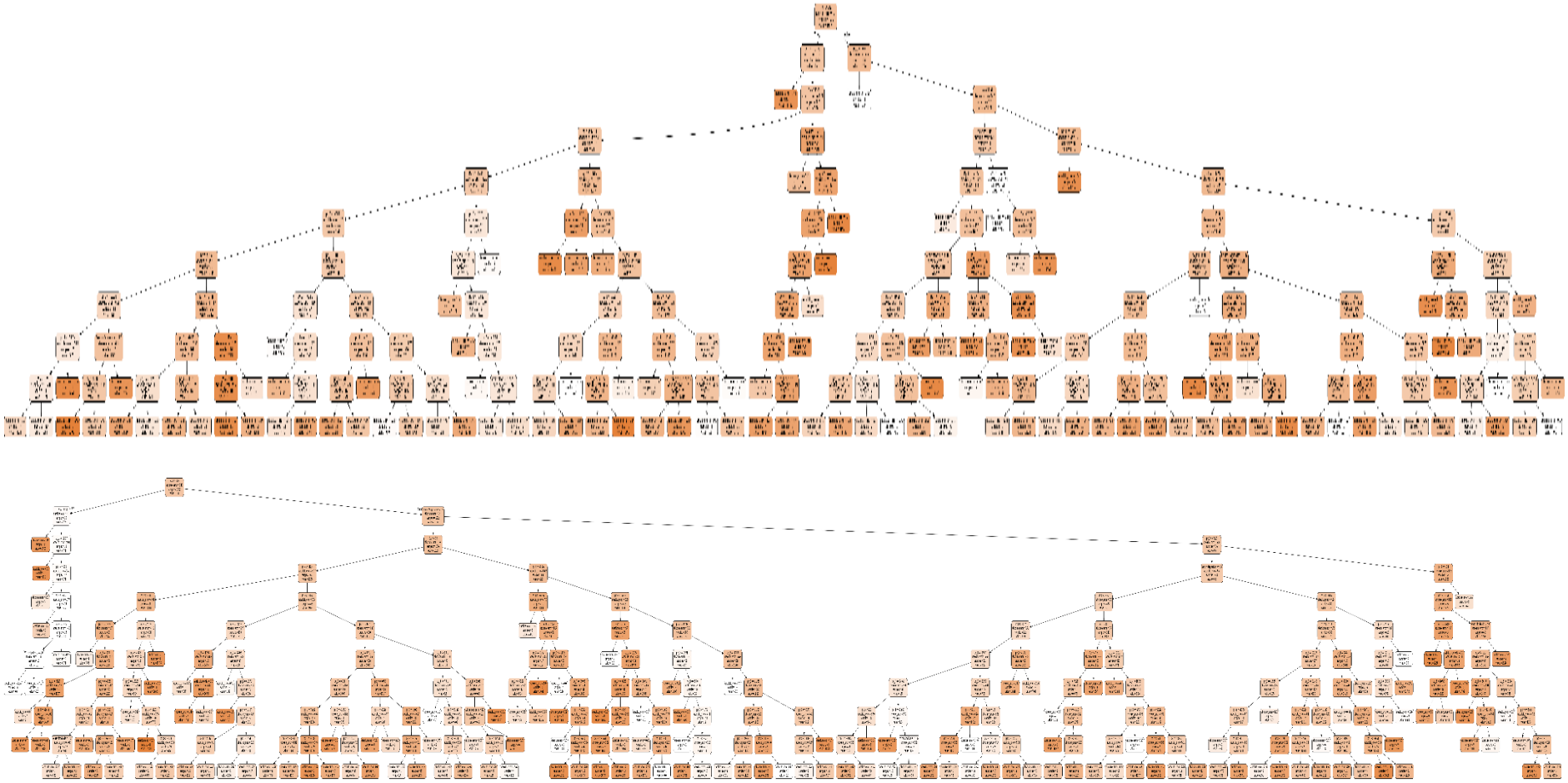
	Race	Serum Magnesium	Sex	White blood cells	pc_1	pc_2	pc_3	pc_4
0	2	0	0	0	-66.933113	-27.421492	56.985991	-31.040740
1	2	0	1	0	69.633384	-22.870550	-12.548636	-22.243195
2	2	1	1	0	-3.252787	-23.962776	-12.613023	-54.840503
3	2	1	0	2	-31.276249	-26.329266	-11.680368	31.495433
4	2	0	1	0	-47.695346	-26.693341	-10.961936	-22.679750

Gambar 4. 47 transformasi pca

pc_1, pc_2, pc_3, dan pc_4 adalah komponen utama yang dihasilkan dari PCA. Nilai-nilai pada komponen utama menunjukkan seberapa besar pengaruh masing-masing individu dalam dataset terhadap komponen tersebut. Nilai positif atau negatif menunjukkan arah kontribusi terhadap komponen utama.

d. Modelling

Model yang digunakan adalah *Random Forest* untuk regresi, dengan menggabungkan hasil dari banyak *decision tree* dan memanfaatkan teknik bootstrap aggregating (bagging), serta mengembangkan beberapa pohon keputusan. *Random Forest* regresi mampu menangkap pola kompleks dalam data serta mengurangi *overfitting*, menjadikannya alat yang efektif dalam mengatasi masalah regresi dengan berbagai jenis data.



Gambar 4. 48 Random Forest

e. Hyperparameter Tunning dan *Evaluation*

Pemodelan *random forest* seringkali mendapatkan model yang *overfiting* atau bahkan model yang tidak memiliki *hyperparameter* optimal untuk mendapatkan nilai maksimal. Masalah dalam menentukan model yang optimal dengan hasil yang maksimal dapat memakan waktu yang lama dan membuang sumber daya bila tujuannya tidak tercapai. Oleh karena itu, metode *hyperparameter tuning* dinilai dapat mengatasi permasalahan tersebut.

Selain itu model akan dievalusi berdasarkan hasil prediksinya terhadap data yang belum pernah dilihat pada *set testing.* Hasil prediksi tersebut akan dinilai berdasarkan MAE (*mean absolute error)* dan metrik evaluasi *c-index*.

Tabel 4. 3 hyperparameter search, mae and c-index

Parameter	Nilai
Bootstrap	*True*
Criterion	*Absolute Error*
Max Depth	10
Min Samples Leaf	10
Number of Estimators	50
MAE	5.58
C-Index	0.5

Penerapan *hyperparameter tunning* menggunakan *grid search* menghasilkan model dengan nilai parameter terbaik yaitu *bootstrap true* digunakan untuk membangun beberapa model dari berbagai subset data yang diambil dengan pengembalian (*sampling* dengan *replacement*). Lalu dengan *criterion* yang dipilih *Absolute Error* dimana itu adalah metrik yang digunakan untuk mengukur akurasi model regresi dengan

menghitung perbedaan absolut antara nilai yang diprediksi dan nilai aktual.

Selanjutnya didapatkan nilai *max depth* 10 yang berarti bahwa batas maksimum kedalaman pohon adalah 10, dimana setiap pohon keputusan yang dibangun dalam model tidak akan memiliki lebih dari 10 level atau lapisan dari akar ke daun. Nilai parameter *Min Samples Leaf* didapatkan sebesar 10, ini berarti bahwa setiap pembagian (*split*) *node* internal dalam pohon keputusan hanya akan terjadi jika ada minimal 10 sampel yang tersedia di *node* tersebut.

Parameter untuk *Number of Estimators* didapatkan sebesar 50, ini menunjukan model akan membangun dan menggabungkan hasil dari 50 pohon dari keputusan individual untuk menghasilkan prediksi akhirnya. Dimana Setiap pohon keputusan memberikan prediksinya sendiri.

Metrik evaluasi MAE didapatkan sebesar 5.58 yang berarti bahwa Setiap prediksi yang dihasilkan oleh model memiliki kesalahan rata-rata sebesar 5.37 dari nilai sebenarnya. Selain itu metrik evaluasi *c-index* menghasilkan nilai sebesar 0.5. *C-index* menilai seberapa baik model dapat membedakan antara pasangan data yang memiliki urutan nilai berbeda.

f. Testing/Deployment

Berdasarkan hasil performa maka akan diuji coba langsung melalui *library streamlit* untuk memprediksi waktu kelangsungan hidup pasien berdasarkan data masukan.

Gambar 4. 49 deployment

4.2 Pembahasan

Prediksi waktu kelangsungan hidup pasien penderita *cardiovascular* dengan metode *random forest* belum bisa memberikan performa yang cukup baik dalam menentukan waktu kelangsungan hidup pasien *cardiovascular*. Ini didasari oleh nilai MAE yaitu 5.37, dimana ini menunjukkan secara rata-rata, prediksi model berbeda sebesar 5.37 dari nilai yang sebenarnya terjadi. Nilai MAE yang lebih rendah menunjukkan bahwa model memiliki kesalahan prediksi yang lebih kecil dan lebih akurat, namun MAE bisa dianggap cukup baik atau buruk tergantung pada skala dan rentang data asli.

C-index yang didapatkan adalah 0.5, dimana nilai *C-Index* berkisar antara 0 hingga 1. Nilai 1 menunjukkan peringkat prediksi yang sempurna dan nilai 0.5 menunjukkan bahwa model tidak lebih baik daripada tebak-tebakan acak. Ini menunjukkan bahwa model tidak memiliki kemampuan untuk merangking prediksi secara akurat. Adapun hasil prediksi model *Random Forest* regresi untuk beberapa sampel sebagai berikut.

	Race	Serum Magnesium	Sex	White blood cells	pc_1	pc_2	pc_3	pc_4	y_true	prediksi_best_rf
0	2	1	0	0	-253.724184	46.185466	-18.795761	-350.746525	10.471461	8.875612
1	2	1	1	0	-253.892933	46.804720	-18.814668	-350.613306	3.807534	8.986208
2	1	0	1	0	-253.803088	45.843325	-18.779162	-350.649620	15.275799	9.917887
3	1	0	1	0	-253.852185	46.394091	-18.780804	-350.755655	2.239041	9.917887
4	1	2	1	0	-253.756730	46.314019	-18.790424	-350.786738	17.661187	9.221735

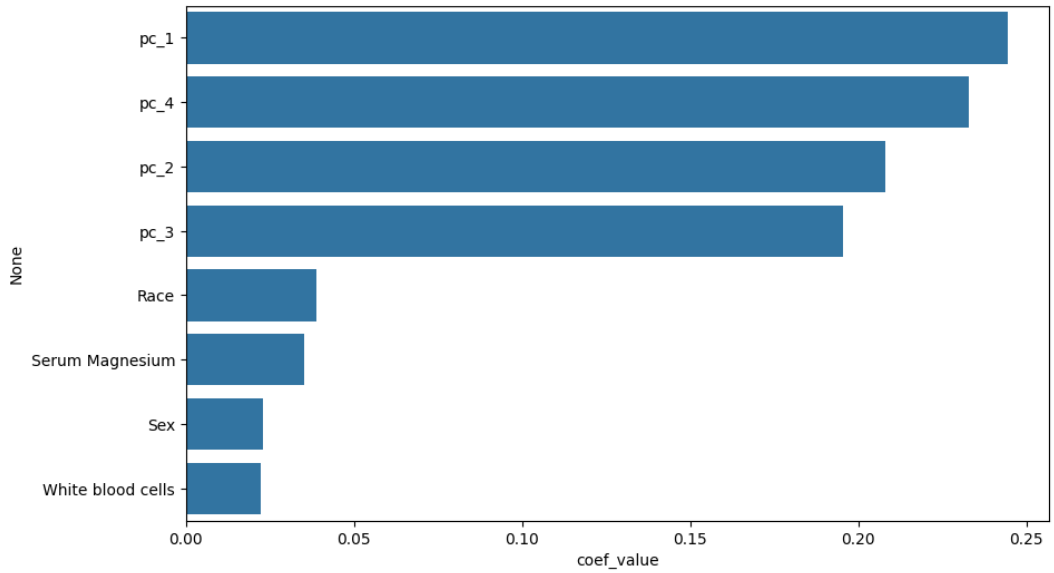
Gambar 4. 50 hasil prediksi

Nilai prediksi (prediksi*_best_rf*) dibandingkan dengan nilai sebenarnya (*y_true*) menunjukkan perbedaan antara hasil prediksi dan observasi. Nilai prediksi cenderung berada dalam rentang yang lebih sempit dibandingkan dengan nilai sebenarnya. Hal ini bisa menunjukkan bahwa model cenderung menghasilkan prediksi yang tidak memiliki variablitas yang cukup, sehingga mengindikasikan *underfitting*.

Hal ini disebabkan karena data awal menunjukkan distribusi yang belum merata atau yang belum terdistribusi dengan baik, yang mengakibatkan model mengalami kesulitan dalam mempelajari pola yang konsisten. Ketika data tidak terdistribusi dengan baik, model mungkin kesulitan dalam menemukan hubungan yang berarti antara fitur dan variable target, sehingga mengurangi kemampuan model untuk memprediksi.

Adapun keberadaan *outlier* yang sangat mempengaruhi kualitas model dan seringkali menyebabkan *overfitting,* sehingga mengakibatkan penurunan performa pada data baru. Dengan menghapus *outlier* jumlah data berkurang secara signifikan dan mengakibatkan hilangnya informasi yang berharga. Penanganan *outlier* dengan imputasi pun menyebabkan data tidak terdistribusi dengan baik, sehingga mengurangi kualitas model dan akurasi prediksi. Hal ini yang mengakibatkan model tidak mampu mengidentifikasi pola yang benar atau relevan data.

Untuk mengetahui pentingnya setiap fitur dalam model prediksi maka perlu dilakukan pencarian fitur terbaik sebagai berikut.



Gambar 4. 51 best feature importance

Ini menunjukkan bahwa firut pc_1 dan pc_4 memiliki koefisien cukup tinggi yang berarti sangat penting dalam model dan memiliki pengaruh terbesar dalam model prediksi, serta memberikan kontribusi terhadap prediksi cukup besar. Selanjutnya pc_2 dan pc_3 kedua komponen ini juga memiliki nilai koefisien yang signifikan meskipun sedikit lebih rendah, ini menunjukkan bahwa variasi yang ditangkap juga relevan untuk prediksi.

Sedangkan untuk fitur lainnya juga tetap berpengaruh dalam model, meskipun kontribusinya lebih kecil. Fitur-fitur ini memiliki dampak prediksi model, tetapi kurang signifikan dibandingkan fitur-fitur PC.