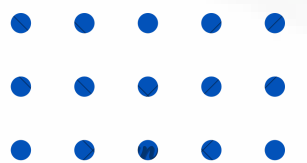


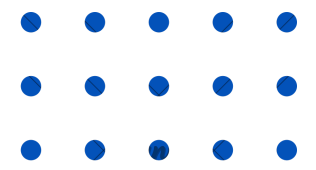
# **Penerapan Algoritma K-Nearest Neighbors dan Generic Algorithm untuk Prediksi Diagnosis Penyakit Jantung**

Oleh :

Muhammad Ragil / 103012300015  
Raditya Bagas Argana / 103012300205  
Luhung Setyo Pambudi / 103012330045



# Pendahuluan

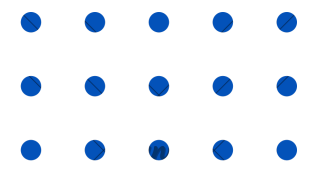


## Latar Belakang

- **Urgensi Medis:** Penyakit jantung tetap menjadi penyebab utama kematian global; kecepatan diagnosis sangat krusial.
- **Kompleksitas Data:** Diagnosis melibatkan banyak variabel klinis (usia, tekanan darah, enzim jantung seperti Troponin & KCM).
- **Solusi Teknologi:** Pemanfaatan Machine Learning sebagai sistem pendukung keputusan medis yang objektif.
- **Metode:** Penggabungan KNN untuk klasifikasi dan Genetic Algorithm (GA) untuk optimasi model.

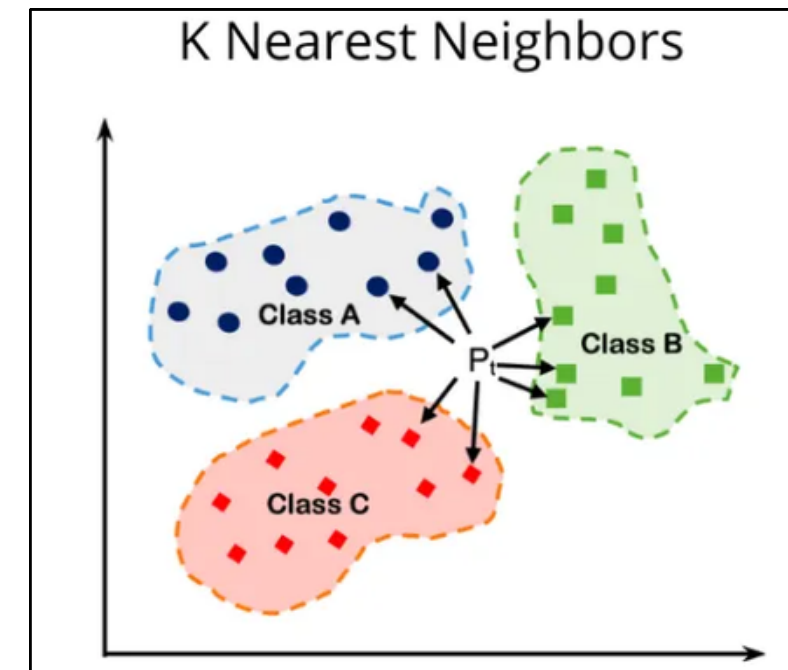


# Pendahuluan



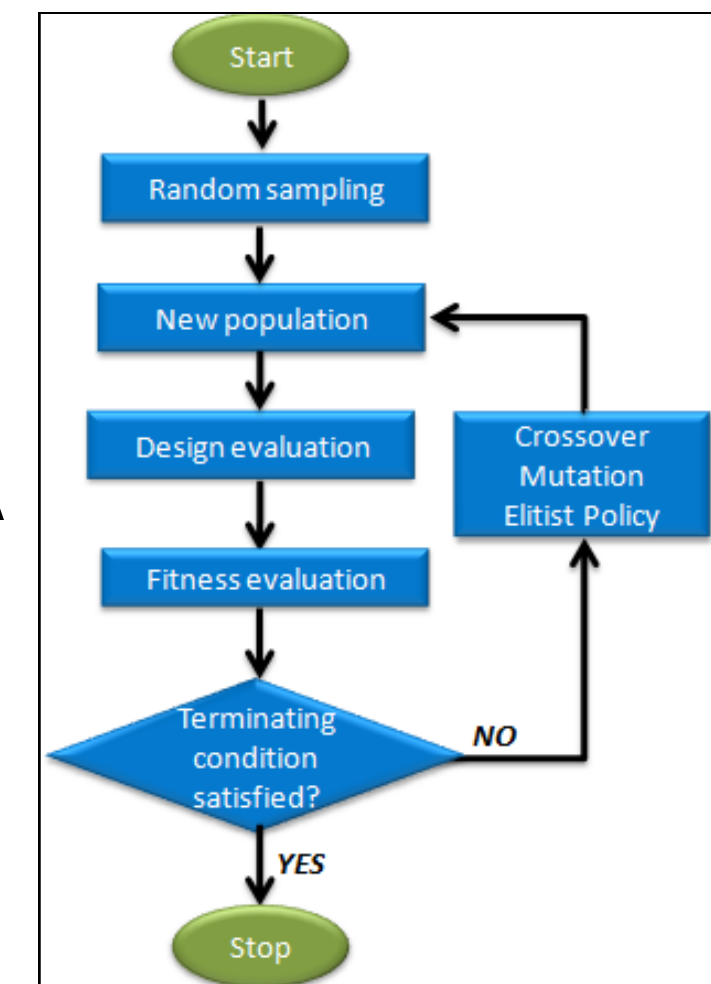
## Memahami Algoritma

- **K-Nearest Neighbors (KNN):** Klasifikasi berdasarkan kedekatan jarak antar data (pola medis yang serupa).
- **Genetic Algorithm (GA):** Digunakan untuk mencari parameter atau fitur terbaik untuk model KNN

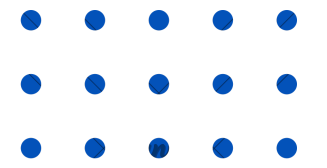


Visual Algoritma KNN

Flow Algoritma GA



# Pendahuluan



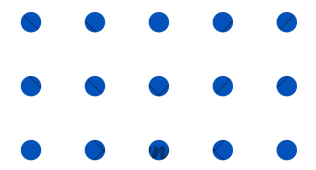
## Batasan Masalah

Untuk menjaga fokus penelitian, batasan yang ditetapkan adalah:

- **Data:** Menggunakan Heart Attack Classification Training Dataset (1.319 baris).
- **Fitur Medis:** Fokus pada 8 variabel (usia, gender, impuls, tekanan darah, glukosa, KCM, troponin).
- **Output:** Klasifikasi biner (Positive atau Negative).
- **Teknologi:** Implementasi menggunakan ekosistem Python (Pandas, Scikit-Learn, Matplotlib).



# Pendahuluan

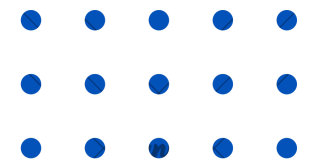


## Tujuan Penelitian

- **Analisis Data:** Melakukan Exploratory Data Analysis (EDA) untuk melihat korelasi fitur medis.
- **Implementasi KNN:** Membangun model prediksi serangan jantung yang tangguh.
- **Optimasi GA:** Mengintegrasikan Algoritma Genetika untuk meningkatkan performa model.
- **Evaluasi:** Mengukur tingkat akurasi untuk memastikan kelayakan model sebagai alat bantu medis.



# Paparan Data



Data yang diambil berasal dari heart attack classification training dataset. Dataset ini memiliki jumlah 1319 baris data dan terdiri dari 9 kolom variabel parameter medis. Variabel-variabel tersebut meliputi age ,gender, impluse, pressurehight, pressurelow, glucose, kcm , dan troponin. Lalu di kolom terakhir ada class sebagai label diagnosis ( positif dan negatif ).

	age	gender	impluse	pressurehight	pressurelow	glucose	kcm	troponin	class
0	62	1	66	160	83	160.0	1.80	0.012	negative
1	21	1	96	98	46	296.0	6.75	1.060	positive
2	55	1	64	160	77	270.0	1.99	0.003	negative
3	64	1	70	120	55	270.0	13.87	0.122	positive
4	55	1	64	112	65	300.0	1.08	0.003	negative
...	...	...	...	...	...	...	...	...	...
1314	44	1	94	122	67	204.0	1.63	0.006	negative
1315	66	1	84	125	55	149.0	1.33	0.172	positive
1316	45	1	85	168	104	96.0	1.24	4.250	positive
1317	54	1	58	117	68	443.0	5.80	0.359	positive
1318	51	1	94	157	79	134.0	50.89	1.770	positive

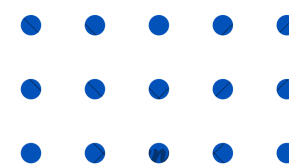
1319 rows x 9 columns

Berikut adalah keterangan untuk dataset Heart Attack Classification

- Age : Umur pasien (tahun)
- Gender : Jenis kelamin (biasanya 1 = Laki-laki, 0 = Perempuan)
- Impluse : Frekuensi denyut jantung (bpm)
- Pressurehight : Tekanan darah sistolik (mmHg)
- Pressurelow : Tekanan darah diastolik (mmHg)
- Glucose : Tingkat glukosa dalam darah
- Kcm : Konsentrasi enzim CK-MB (Creatine Kinase-MB)
- Troponin : Konsentrasi protein Troponin dalam darah
- Class : Diagnosis akhir (positive jika serangan jantung dan negative jika tidak)

Sumber : <https://www.kaggle.com/code/mragpavank/pima-indians-diabetes-database>

# Statistik Data

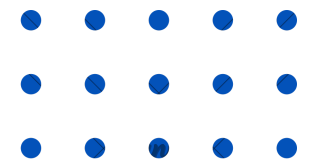
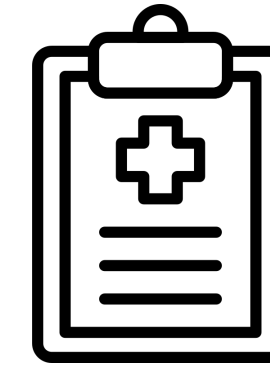


Dengan menggunakan fungsi dari library python yang ada, kita dapat melihat gambaran umum data melalui statistik yang ada

	count	mean	std	min	25%	50%	75%	max
age	1319.0	56.190296	13.646558	14.000	47.000	58.000	65.0000	103.0
gender	1319.0	0.659591	0.474027	0.000	0.000	1.000	1.0000	1.0
impluse	1319.0	78.338135	51.630760	20.000	64.000	74.000	85.0000	1111.0
pressurehigh	1319.0	127.170584	26.122720	42.000	110.000	124.000	143.0000	223.0
pressurelow	1319.0	72.269143	14.033924	38.000	62.000	72.000	81.0000	154.0
glucose	1319.0	146.634344	74.923045	35.000	98.000	116.000	169.5000	541.0
kcm	1319.0	15.274306	46.327083	0.321	1.655	2.850	5.8050	300.0
troponin	1319.0	0.360942	1.154568	0.001	0.006	0.014	0.0855	10.3

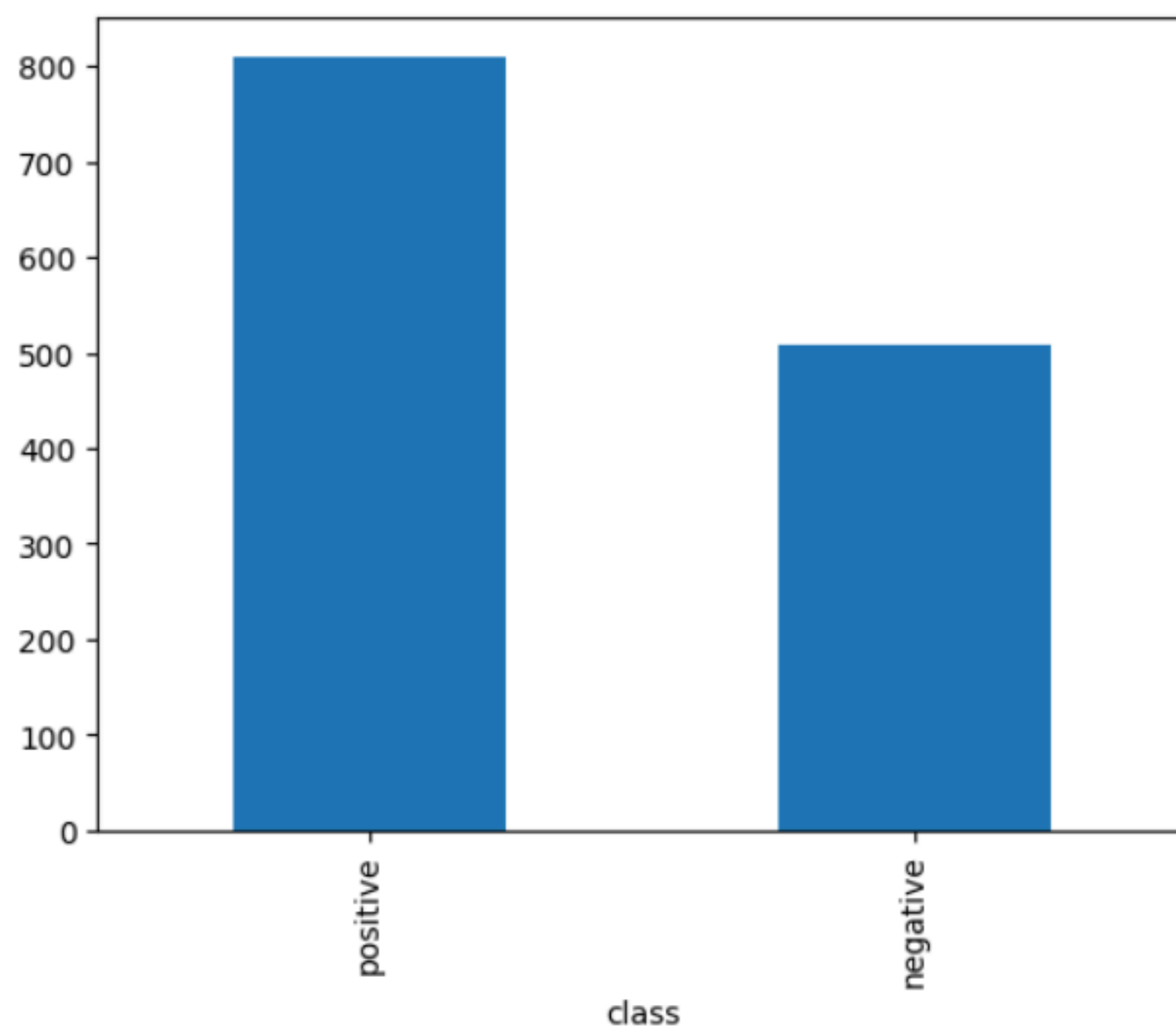
- **Count:** Memastikan jumlah observasi pada setiap kolom tetap konsisten, yaitu sebanyak 1.319 baris data.
- **Mean & Std:** Melihat nilai rata-rata dan standar deviasi untuk mengukur penyebaran atau variansi data.
- **Min & Max:** Mengidentifikasi nilai terendah dan tertinggi guna mendeteksi adanya potensi outlier atau pencilan yang ekstrem.
- **Quartiles (25%, 50%, 75%):** Memahami persebaran data pada rentang tertentu, di mana nilai 50% (median) menunjukkan titik tengah dari data tersebut.

# VISUALISASI VARIABEL DIAGNOSIS



```
df['class'].value_counts().plot(kind='bar')
```

<Axes: xlabel='class'>



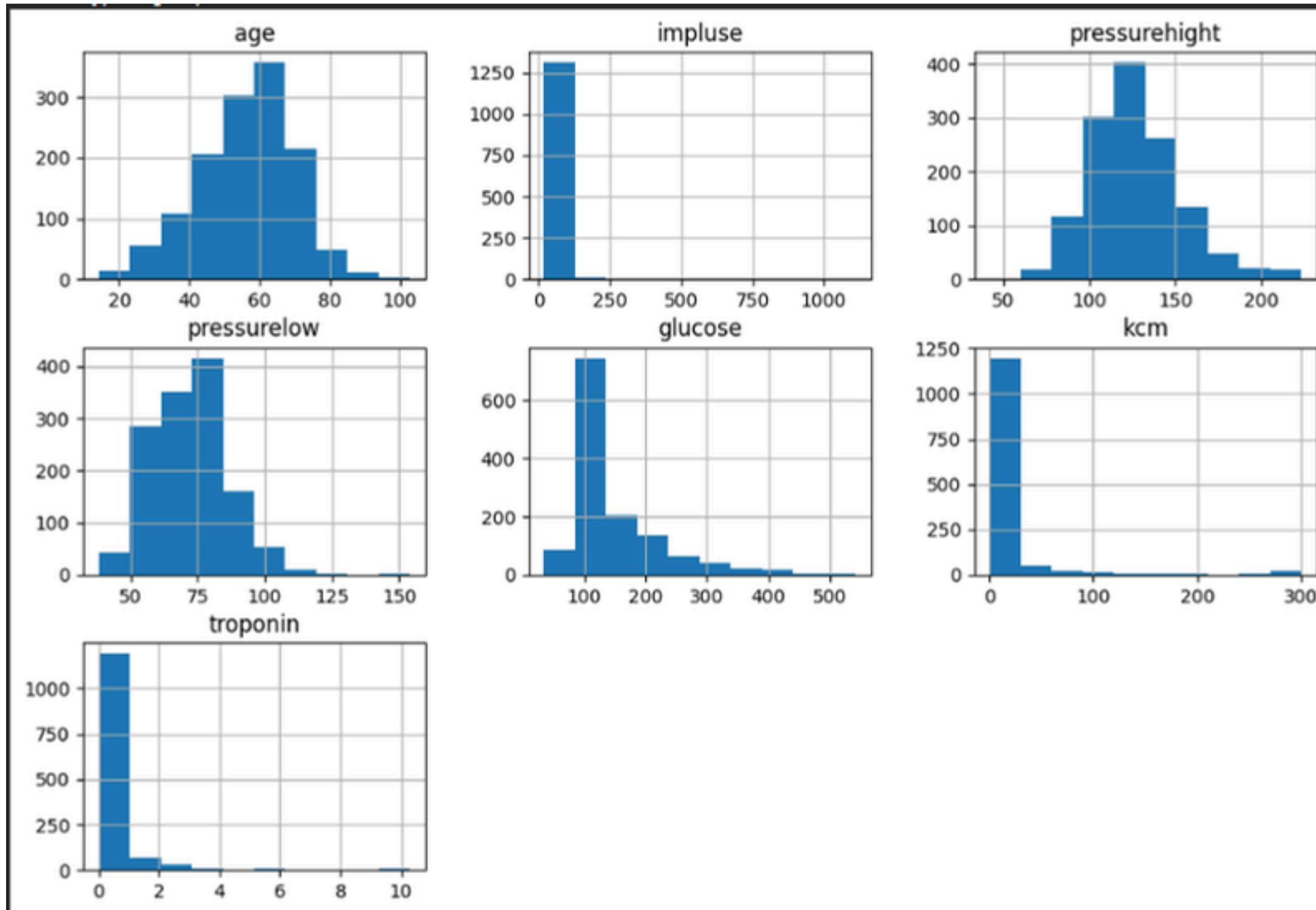
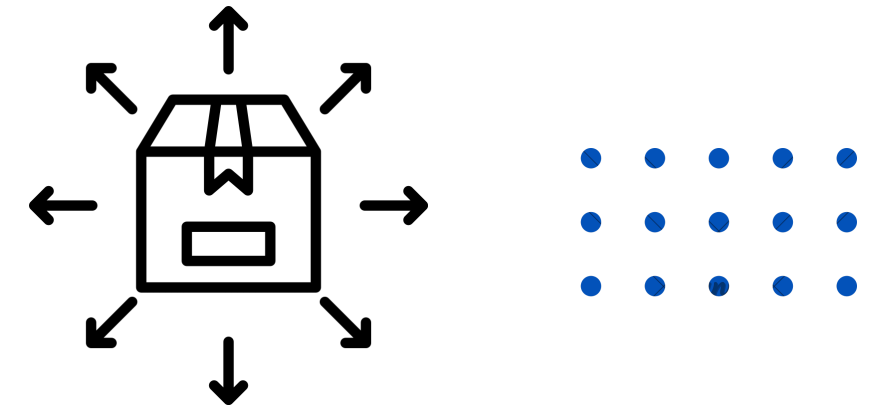
## Tujuan Visualisasi

Visualisasi ini bertujuan mengevaluasi distribusi data untuk mendeteksi potensi ketidakseimbangan kelas (class imbalance).

## Analisis Grafik

Berdasarkan grafik yang dihasilkan, terlihat bahwa kategori diagnosis positif memiliki frekuensi lebih tinggi dibanding negatif. Meski demikian, selisih antar kelas tidak ekstrem, sehingga dataset tetap dinilai layak dan representatif untuk proses klasifikasi selanjutnya.

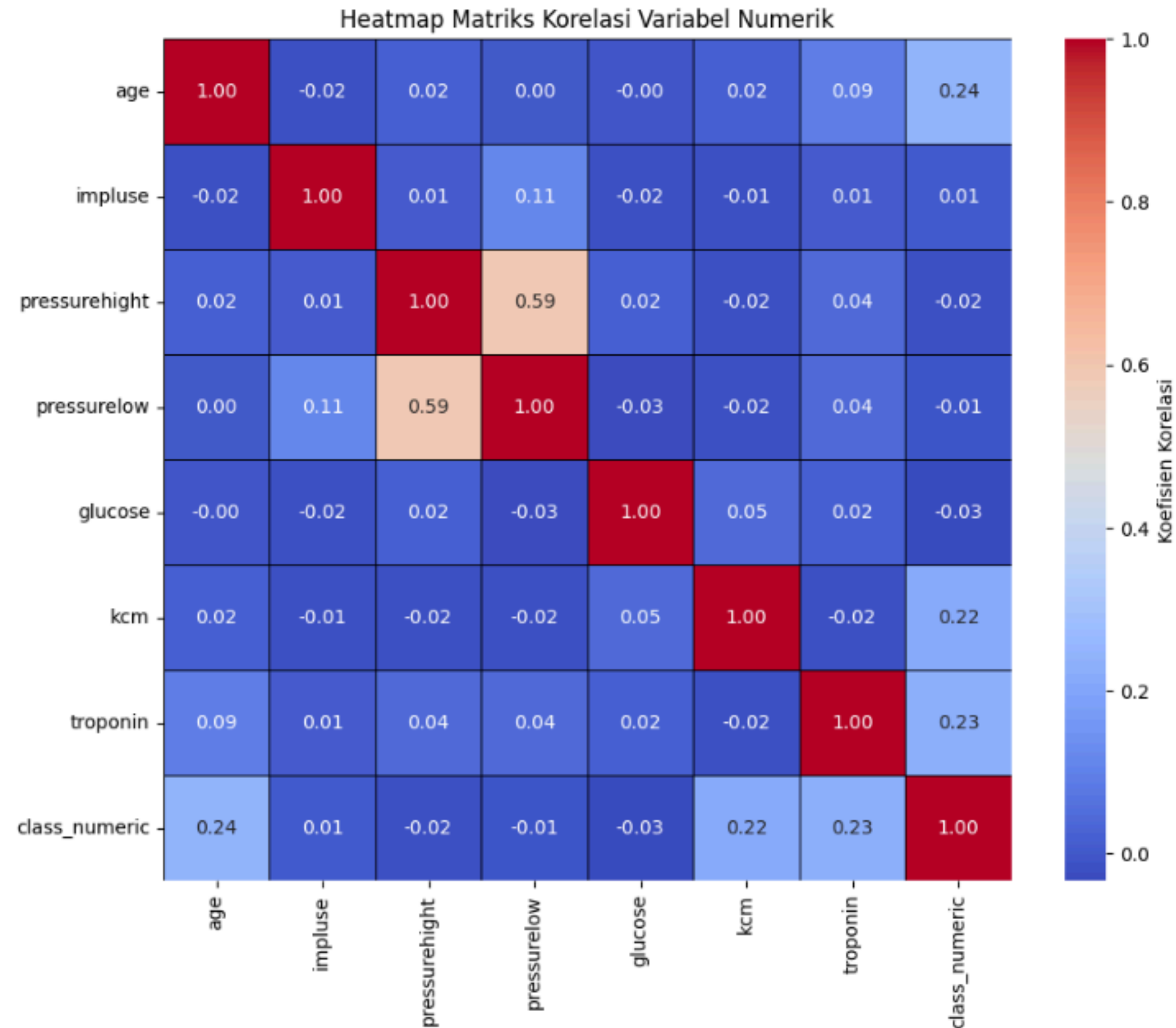
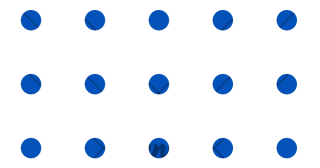
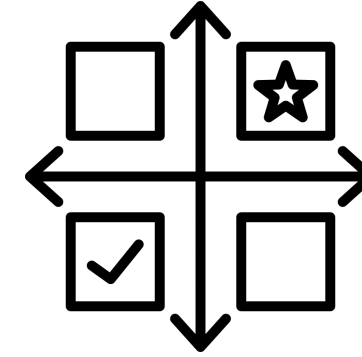
# VISUALISASI PENYEBARAN DATA FITUR



Pada tahap ini bertujuan untuk melihat persebaran data pada setiap fitur, menggunakan histogram untuk mengidentifikasi tiga aspek utama data:

- Kerapatan: Menentukan titik konsentrasi nilai tertinggi pada setiap fitur.
- Skewness: Menilai simetri distribusi (Normal pada Age/Pressure, Miring/Skewed pada KCM/Troponin).
- Anomali: Mendeteksi pencilan (outliers) dan variabel dengan sebaran yang terlalu sempit secara visual.

# VISUALISASI MATRIKS KORELASI



## Tujuan Visualisasi

Visualisasi matriks korelasi menggunakan heatmap ini bertujuan untuk memetakan hubungan linier antar variabel medis serta pengaruhnya terhadap target diagnosis secara menyeluruh.

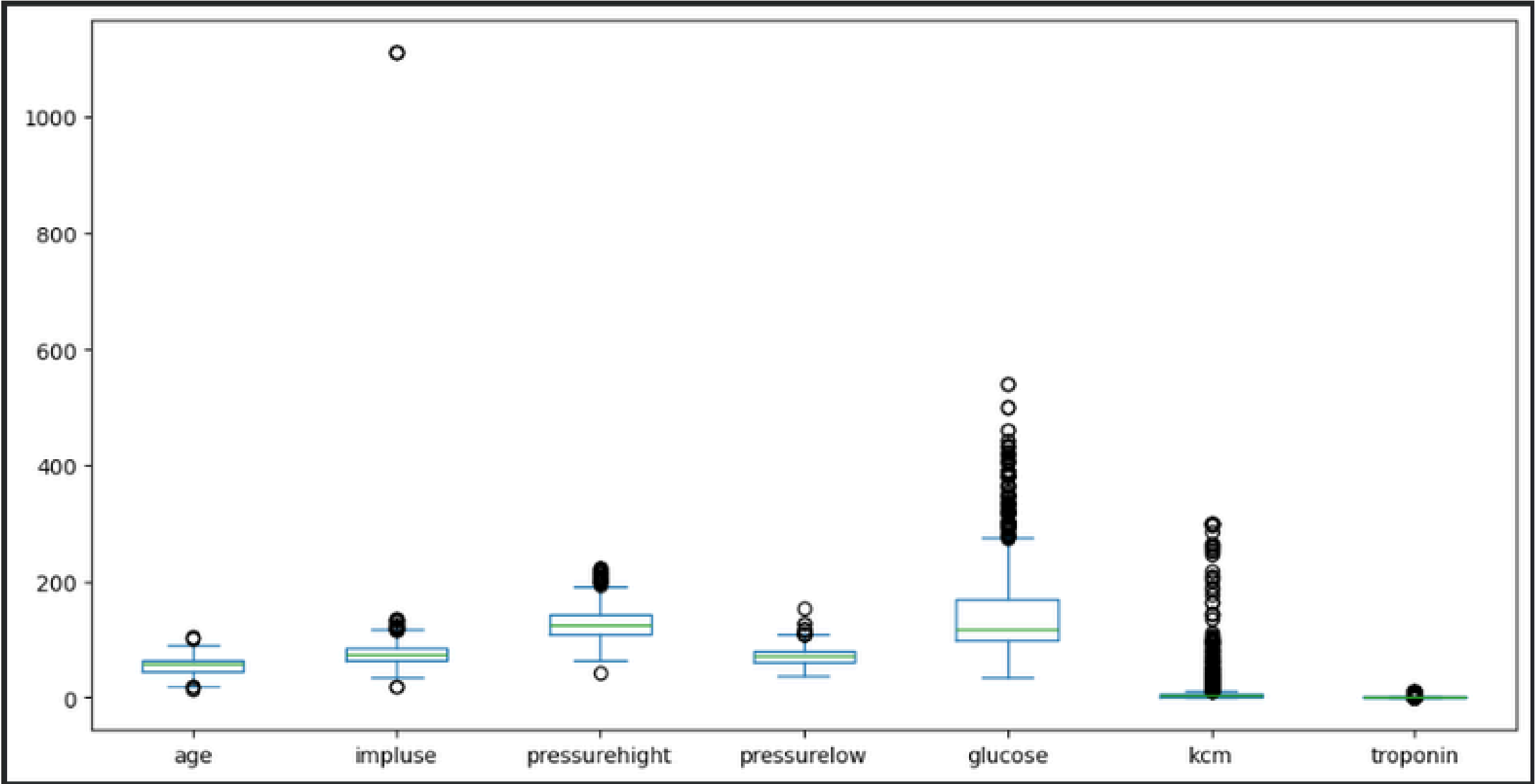
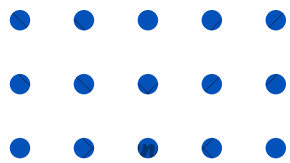
## Analisis Matriks Korelasi

Berdasarkan matriks korelasi, terdapat 3 fitur yang memiliki korelasi paling baik terhadap class\_numeric (diagnosis jantung). Hal ini dapat dilihat dari nilai baris tertinggi pada heatmap, yaitu:

Fitur	Age	Troponin	KCM
Nilai Korelasi	0,24	0,23	0,22

Sementara itu, variabel lain hanya memiliki nilai korelasi dengan class\_numeric (diagnosis jantung) yang mendekati nilai 0 atau bahkan bernilai negatif.

# VISUALISASI BOX PLOT



**Tujuan Visualisasi**  
Memetakan distribusi data untuk mendeteksi apakah outlier merupakan variasi medis yang valid atau kesalahan input (input error).

Analisis Fitur Utama		
Fitur	Karakteristik Distribusi	Implikasi pada Model (KNN)
Age	Stabil & terpusat; minim outlier.	Sangat efektif untuk perhitungan jarak yang konsisten.
Troponin	Rentang sempit; tanpa outlier ekstrem.	Indikator medis yang sangat informatif meski variasi kecil.
KCM	Asimetris; banyak outlier bernilai tinggi.	Perlu penanganan khusus agar tidak mendominasi bobot jarak.