

Eksplorasi Model Machine Learning *Heart Failure Prediction Dataset*

[Heart Failure Prediction Dataset \(kaggle.com\)](https://www.kaggle.com/datasets/lucalmeida/heart-failure-prediction)

Gagal jantung adalah suatu kondisi medis serius yang erat kaitannya dengan penyakit kardiovaskular, yang saat ini menjadi penyebab kematian utama di seluruh dunia. Penyakit kardiovaskular adalah penyebab sekitar 17,9 juta kematian setiap tahun, yang mewakili sekitar 31% dari total kematian global. Dari seluruh kematian yang disebabkan oleh penyakit kardiovaskular, empat dari lima kasusnya disebabkan oleh serangan jantung dan stroke, dengan sekitar sepertiga dari kematian ini terjadi pada usia di bawah 70 tahun.

Gagal jantung merupakan komplikasi umum yang terkait erat dengan penyakit kardiovaskular. Untuk memahami dan mengelola kondisi ini lebih baik, data dan informasi sangat penting. Inilah sebabnya mengapa dataset yang berisi 11 fitur penting telah dikembangkan untuk membantu dalam memprediksi risiko terkena penyakit jantung.

Individu yang memiliki penyakit kardiovaskular atau berisiko tinggi (karena memiliki satu atau lebih faktor risiko seperti hipertensi, diabetes, hiperlipidemia, atau riwayat penyakit lainnya) memerlukan deteksi dini dan manajemen yang tepat. Dalam konteks ini, teknologi machine learning dapat memainkan peran yang sangat berarti. Model machine learning dapat membantu mengidentifikasi individu yang berisiko tinggi, memperkirakan kemungkinan terkena penyakit jantung, dan bahkan membantu mengembangkan strategi pengelolaan yang lebih efektif. Dengan pendekatan ini, kita dapat lebih efisien dalam mencegah dan mengatasi penyakit kardiovaskular, serta memperpanjang umur dan meningkatkan kualitas hidup bagi banyak individu.

Berikut merupakan tabel atribut dan keterangan dari dataset *heart failure prediction*

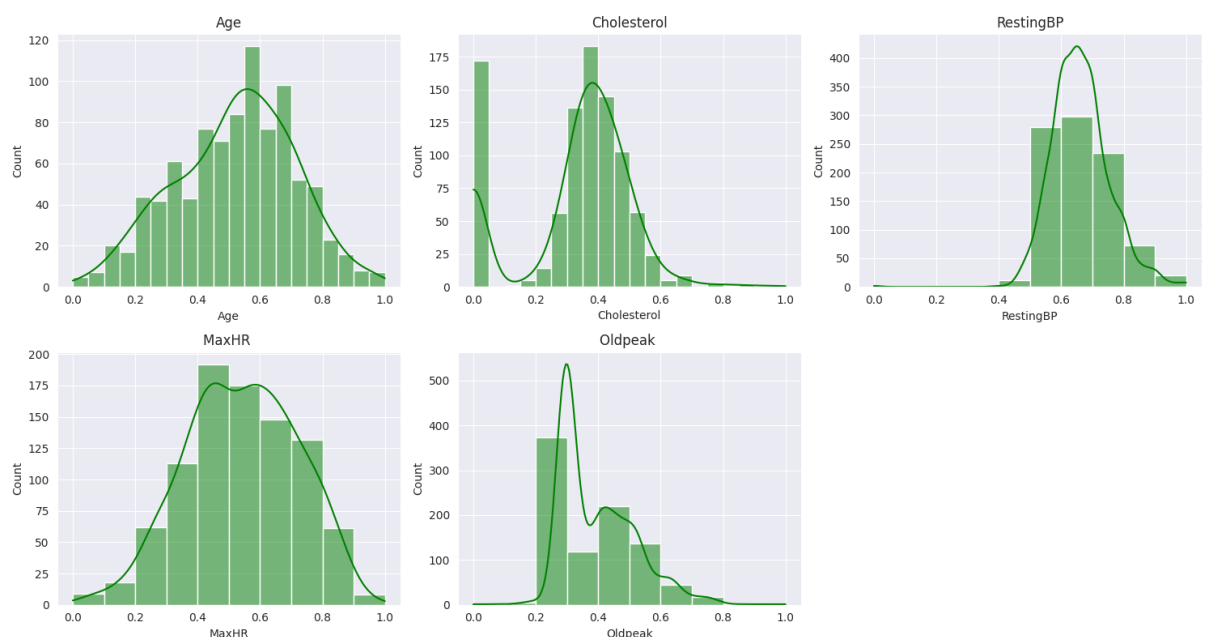
Atribut	Deskripsi	Nilai
Age	Usia pasien	[tahun]
Sex	Jenis kelamin pasien	[M: Laki-laki, F: Perempuan]
ChestPainType	Jenis nyeri dada	[TA: Angina Tipikal, ATA: Angina Atipikal, NAP: Nyeri Non-Anginal, ASY: Asimtomatik]
RestingBP	Tekanan darah istirahat	[mm Hg]
Cholesterol	Kolesterol serum	[mm/dl]
FastingBS	Gula darah saat puasa	[1: jika FastingBS > 120 mg/dl, 0: jika sebaliknya]
RestingECG	Hasil elektrokardiogram istirahat	[Normal: Normal, ST: Adanya abnormalitas gelombang ST-T (inversi gelombang T dan/atau elevasi atau depresi ST > 0.05 mV), LVH: Kemungkinan atau pasti adanya hipertrofi ventrikel kiri menurut kriteria Estes]
MaxHR	Denyut jantung maksimal yang dicapai	[Nilai numerik antara 60 dan 202]

Atribut	Deskripsi	Nilai
ExerciseAngina	Angina yang dipicu oleh latihan	[Y: Ya, N: Tidak]
Oldpeak	Oldpeak = ST	[Nilai numerik yang diukur dalam depresi]
ST_Slope	Kemiringan segmen ST saat puncak latihan	[Up: menaik, Flat: datar, Down: menurun]
HeartDisease	Kelas keluaran	[1: penyakit jantung, 0: Normal]

A. Exploratory Data Analysis.

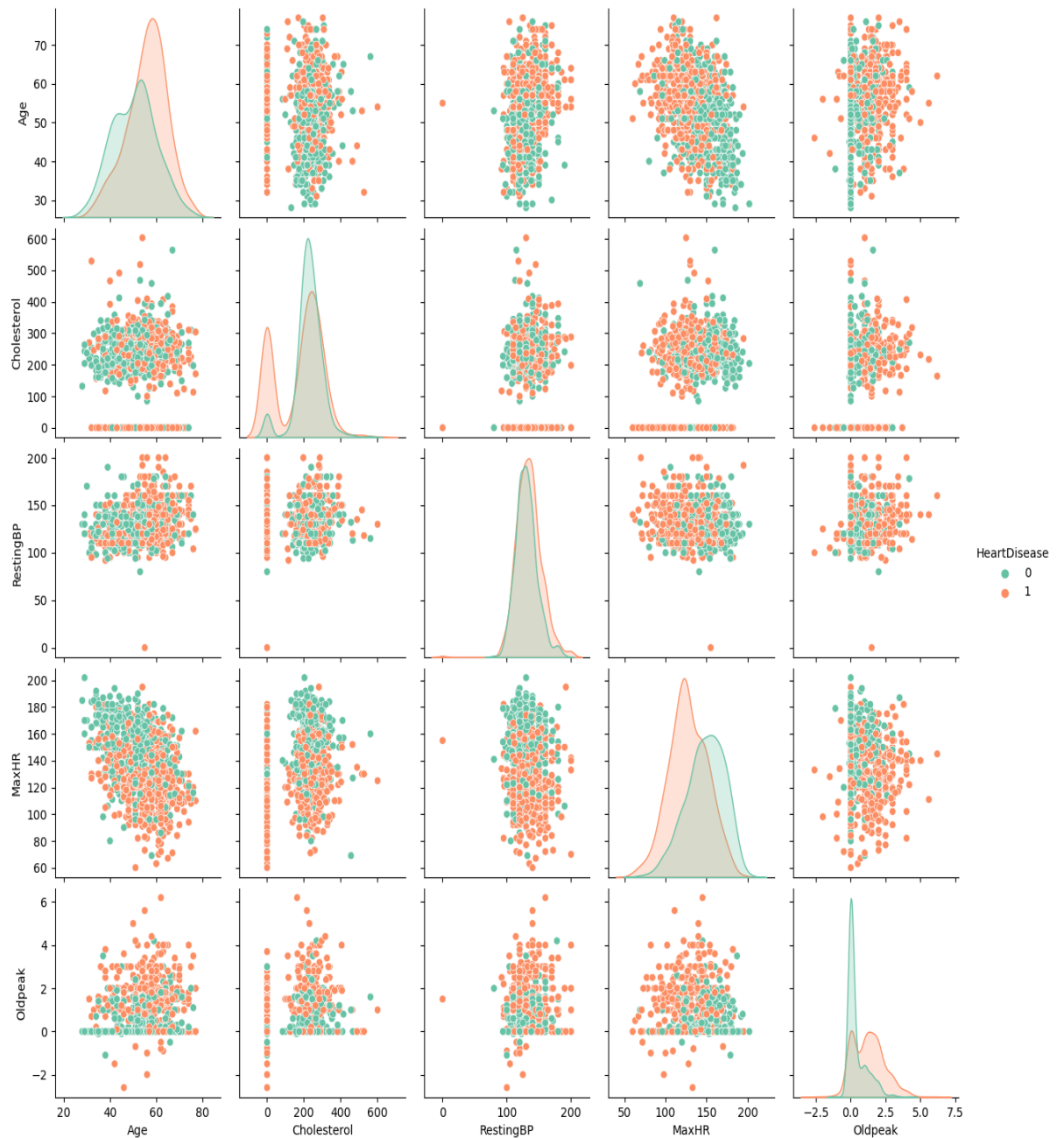
EDA adalah proses investigasi awal dalam analisis data yang bertujuan untuk memahami struktur dan karakteristik data, mengidentifikasi pola dan anomali, serta mendapatkan wawasan awal sebelum melakukan analisis statistik atau pemodelan data yang lebih mendalam. Tujuan EDA adalah untuk membantu analisis data dalam merumuskan pertanyaan analisis, mengungkapkan informasi yang mungkin tersembunyi dalam data, dan memahami konteks data dengan menggunakan berbagai teknik visualisasi dan eksplorasi data.

Gambar 1. Grafik Histogram untuk data numerik



Histogram adalah visualisasi grafis yang digunakan untuk menampilkan distribusi data numerik dan memungkinkan kita untuk melihat sebaran nilai-nilai dalam data dan memahami pola distribusi, termasuk frekuensi kemunculan nilai-nilai tertentu. Berdasarkan grafik histogram pada gambar 1, sebaran data cenderung normal pada data *Age*, *Cholesterol*, *RestingBP*, dan *MaxHR*. Sedangkan, pada data *Oldpeak* persebarannya cenderung miring/skewed.

Gambar 2. Scatter plot untuk data numerik



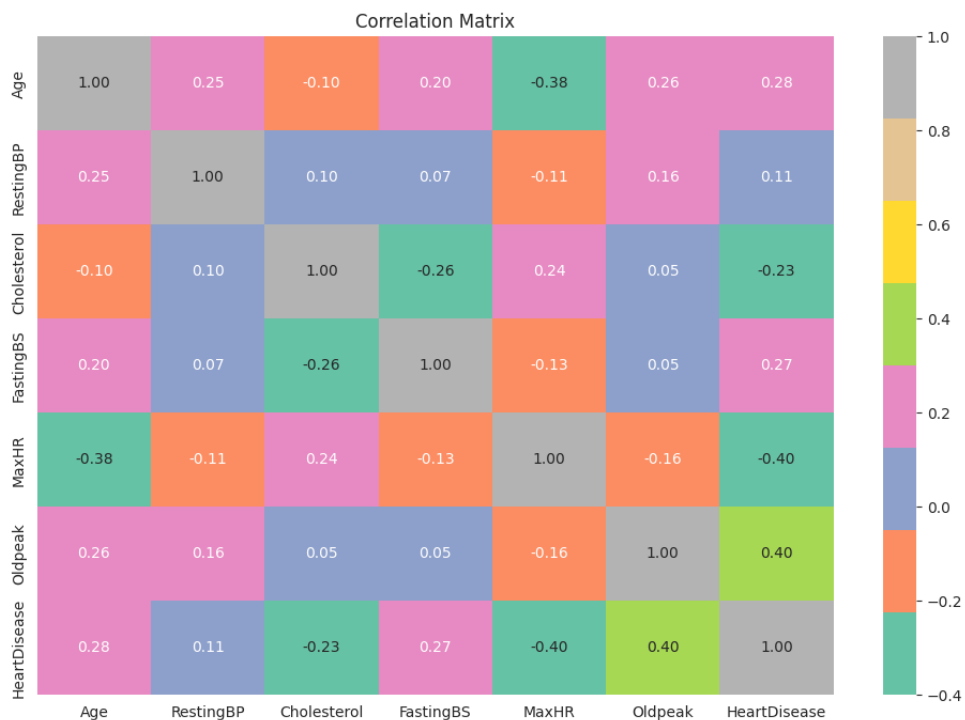
Gambar 2 menunjukkan *scatter plot* untuk data numerik. *Scatter plot* adalah visualisasi data yang dapat membantu dalam mengidentifikasi korelasi, pola, atau hubungan antar variabel tertentu. Berdasarkan gambar 2, terlihat bahwa hanya pasangan data *Age* vs *MaxHR* yang menunjukkan adanya korelasi berupa linear regression positif. Hal ini mengindikasikan bahwa semakin tinggi tua usia seseorang, semakin tinggi pula denyut jantung maksimalnya. Selain itu, terlihat juga terdapat beberapa outlier yang terlihat pada titik-titik yang jauh dari kelompok lainnya.

Gambar 3. Grafik batang untuk data kategori



Gambar 3 menunjukkan persebaran penderita penyakit yang divisualisasikan dengan batang berwarna orange dan pasien yang sehat yang divisualisasikan dengan warna hijau. Berdasarkan visualisasi ini, terlihat bahwa persebaran data antara yang sehat dan yang sakit cenderung tidak seimbang.

Gambar 4. Correlation Matrix



Gambar 4 menunjukkan matriks korelasi antar fitur dan label. Terlihat dari warna-warna dan nilai yang didapat, fitur-fitur yang ada sudah memiliki korelasi yang cukup baik, terlihat pada nilai-nilainya yang tidak ada kecenderungan mendekati 0. Oleh karena itu, pada eksplorasi kali ini, tidak ada fitur yang dibuang.

B. Data Preparation

Sebelum memasuki tahap pemodelan, langkah yang perlu dilakukan adalah persiapan data. Tujuannya adalah memastikan bahwa data yang digunakan memiliki format yang sesuai dan dapat dianalisis dengan baik. Dalam proses ini, variabel kategori yang umumnya dalam bentuk teks diubah menjadi representasi numerik menggunakan teknik *one hot encoding*. Selain itu, data numerik juga dilakukan penskalaan (*rescaling*) untuk memastikan nilainya berada dalam rentang yang sama, antara 0 dan 1, sehingga tidak ada variabel yang mendominasi.

Langkah selanjutnya adalah memisahkan kolom fitur dari labelnya. Kolom yang akan digunakan untuk prediksi, yaitu kolom fitur, dipisahkan dari kolom label, yaitu kolom "HeartDisease." Setelah pemisahan ini, pada model *supervised learning*, data dibagi menjadi dua set terpisah, yaitu data pelatihan (*training data*) yang digunakan untuk melatih model dan data validasi yang digunakan untuk menguji kinerja model. Pada model *Unsupervised learning* tidak dilakukan pemisahan data karena tidak membutuhkan data label dalam pelatihannya.

C. Model Machine Learning

a. Decision Tree

Decision tree, atau pohon keputusan, adalah model prediktif dalam machine learning yang menggambarkan aliran keputusan berdasarkan serangkaian aturan dan pemilihan fitur. Pohon ini terdiri dari simpul yang mewakili pertanyaan atau aturan, serta cabang-cabang yang menggambarkan kemungkinan jawaban atau keputusan yang mungkin diambil. Pada setiap simpul, algoritma *decision tree* memilih fitur yang paling relevan untuk membagi data menjadi subset yang lebih kecil. Proses pemisahan ini berlanjut hingga mencapai simpul daun, yang mewakili hasil atau prediksi akhir. *Decision tree* sering digunakan dalam klasifikasi (pengelompokan) dan regresi (prediksi nilai berkelanjutan) serta dikenal karena kemampuannya untuk menghasilkan model yang mudah diinterpretasi.

Keuntungan utama dari *decision tree* adalah kemampuan interpretasi yang tinggi, yang memungkinkan pengguna untuk memahami alasan di balik setiap keputusan dan prediksi. Namun, pohon keputusan juga memiliki kelemahan, seperti cenderung *overfitting* pada data pelatihan yang rumit dan sensitif terhadap perubahan dalam data. Untuk mengatasi ini, varian seperti *Random Forest* dan *Extreme Gradient Boosting* (XGBoost) telah dikembangkan untuk meningkatkan performa dan ketahanan terhadap *overfitting*. Tabel 2 menunjukkan performa pelatihan dari *decision tree*, *random forest*, dan XGBoost.

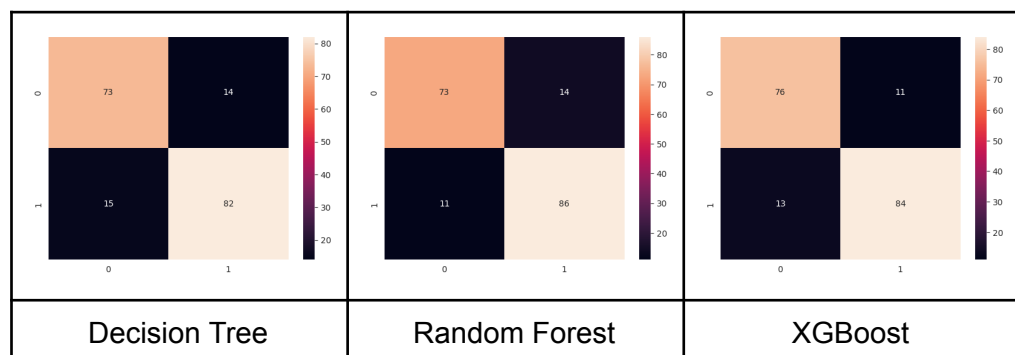
Tabel 2. Performa *decision tree*, *random forest*, dan XGBoost

Model	Akurasi Training	Akurasi Validasi
Decision Tree	90%	84.2%
Random Forest	90.6%	86.4%
XGBoost	97.5%	86.9%

Pada awalnya, ketiga model menunjukkan adanya overfitting yang kuat yang terindikasi pada nilai akurasi training hingga 100%, sementara akurasi validasinya rendah. Oleh karena itu, dilakukan konfigurasi parameter pada masing-masing model untuk mengatasi masalah overfitting. Untuk model Decision Tree dan Random Forest, dilakukan penyesuaian pada parameter seperti `max_depth` dan `min_samples_split`. Sementara untuk model XGBoost, parameter `alpha` disetel agar dapat mengatasi overfitting dan menghasilkan hasil yang lebih konsisten.

Berdasarkan hasil evaluasi akurasi, tiga model yang telah dianalisis menawarkan gambaran yang berbeda tentang performa mereka dalam memprediksi data. Decision Tree, dengan akurasi training sekitar 90%, menunjukkan kemampuannya untuk memahami data pelatihan dengan baik, tetapi akurasi validasi yang lebih rendah sekitar 84.2% mengindikasikan adanya overfitting. Random Forest, dengan akurasi training sekitar 90.6% dan akurasi validasi sekitar 86.4%, menunjukkan kinerja yang lebih baik daripada Decision Tree. Ini menandakan bahwa ensemble learning dalam Random Forest membantu mengurangi overfitting dan menghasilkan hasil yang lebih stabil. Model XGBoost memiliki performa terbaik dengan akurasi training sekitar 97.5% dan akurasi validasi sekitar 86.9%. Hal ini menandakan bahwa XGBoost memiliki kemampuan prediksi yang sangat baik dan relatif tahan terhadap overfitting. Gambar 5 menunjukkan perbandingan dari *confusion matrix* dari ketiga model

Gambar 5 perbandingan *confusion matrix*



b. Artificial Neural Network

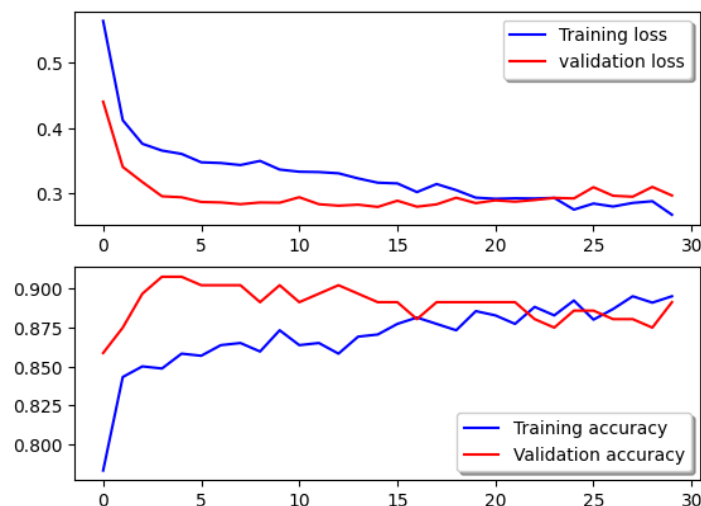
Metode Artificial Neural Network digunakan untuk melakukan klasifikasi pasien berpotensi mengalami penyakit jantung atau tidak. Dalam implementasinya, kita memanfaatkan library TensorFlow untuk membangun model ini. Model ini terdiri dari 5 layer yang berfungsi untuk mencari pola-pola fitur hingga membuat prediksi. Pertama, ada layer Flatten yang berguna untuk meratakan data menjadi bentuk yang lebih sederhana. Dua buah layer Dense, masing-masing dengan 256 unit dan 32 unit, digunakan untuk mengambil representasi fitur-fitur yang relevan dari data. Selanjutnya, ada 1 layer Dropout yang berfungsi untuk mengurangi overfitting dalam model. Terakhir, terdapat layer Dense dengan 1 unit yang bertugas untuk menghasilkan prediksi akhir. Gambar 6 menunjukkan *model summary* dari layer yang dibuat

Gambar 6. *Model summary* ANN

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 15)	0
dense (Dense)	(None, 256)	4096
dense_1 (Dense)	(None, 32)	8224
dropout (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 1)	33
Total params: 12353 (48.25 KB)		
Trainable params: 12353 (48.25 KB)		
Non-trainable params: 0 (0.00 Byte)		

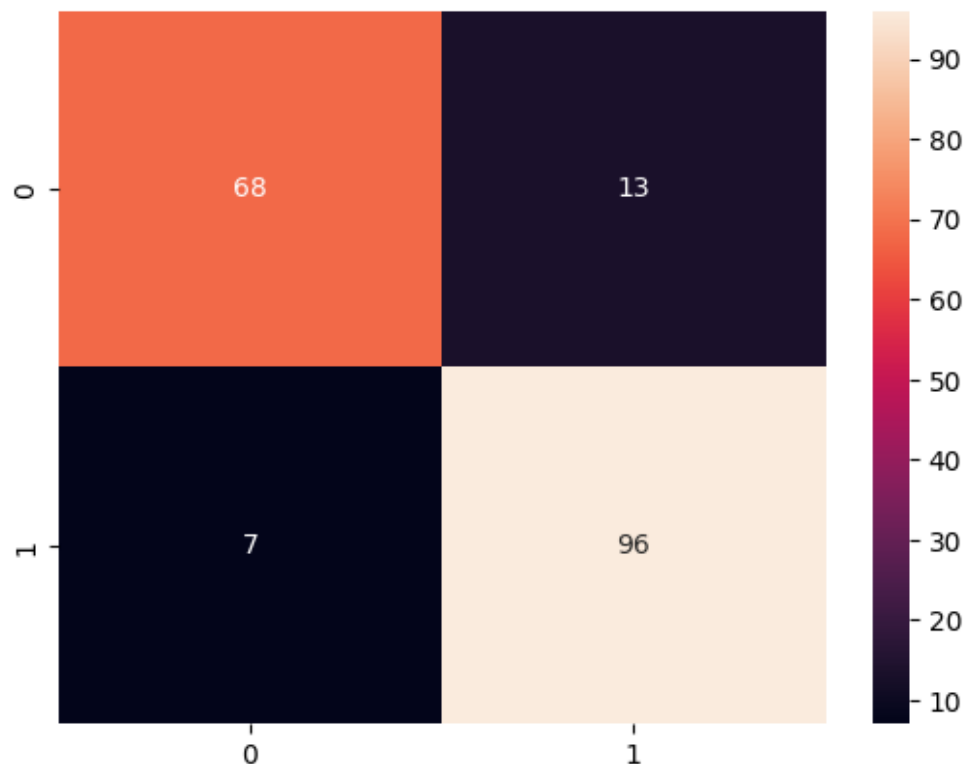
Untuk melatih model, digunakan optimizer *Adam*, *loss function* *binary cross-entropy*, *epoch* sebanyak 30 dan *batch size* sebanyak 32. Gambar 7 menunjukkan grafik akurasi dan loss baik pada data training dan validasi

Gambar 7 grafik akurasi dan loss



Hasil akhir menunjukkan bahwa setelah epoch ke-30, model ANN mencapai akurasi training sebesar 89.5%. Hal ini berarti model mampu memahami dan memprediksi data pelatihan yang telah diberikan dengan baik. Hal yang lebih penting, akurasi validasi yang mencapai 89.1% juga menunjukkan bahwa model mampu melakukan prediksi dengan baik pada data yang belum pernah dilihat sebelumnya. Performa yang tinggi pada kedua metrik ini menunjukkan bahwa model ANN ini memiliki kemampuan yang baik dalam klasifikasi pasien memiliki resiko masalah jantung atau tidak berdasarkan fitur-fitur yang dilatih. Gambar 8 menunjukkan *confusion matrix* dari metode ANN

Gambar 8 *confusion matrix* ANN

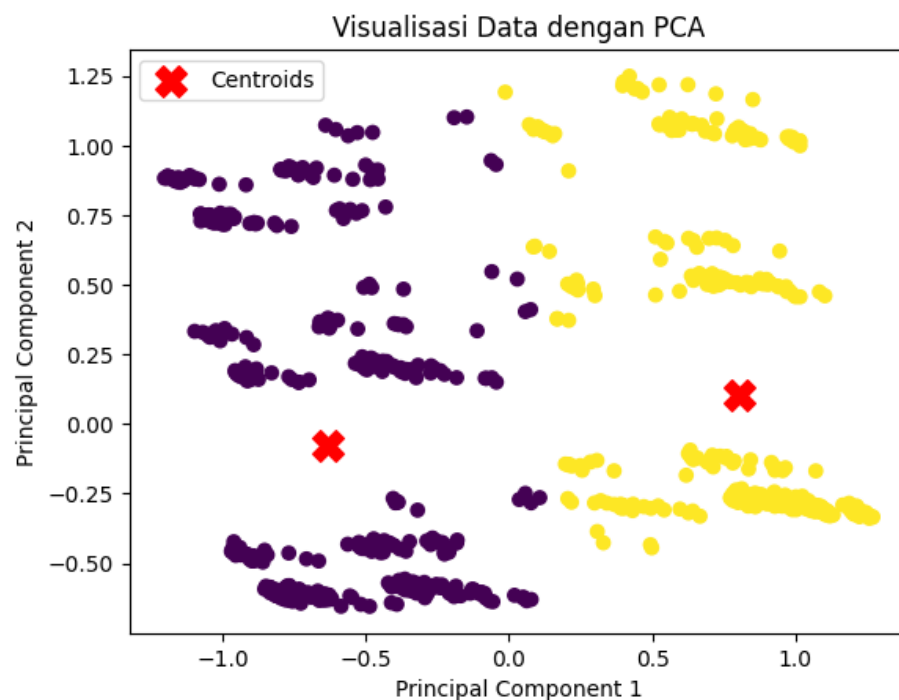


c. *Clustering: K-Means*

K-Means adalah algoritma clustering yang digunakan untuk mengelompokkan kemiripan fitur-fitur tertentu ke dalam kelompok-kelompok yang serupa berdasarkan karakteristiknya. Dalam konteks ini, K-Means digunakan untuk mengidentifikasi kelompok fitur yang memiliki kemiripan dalam hal faktor-faktor risiko yang berkaitan dengan risiko masalah jantung. Dengan membagi pasien ke dalam kelompok-kelompok yang serupa, K-Means dapat membantu dokter atau peneliti dalam pemahaman dan analisis data pasien, serta mengidentifikasi kelompok pasien yang mungkin memiliki risiko yang lebih tinggi atau lebih rendah terhadap masalah jantung.

Untuk membuktikan apakah algoritma cocok atau tidak dalam hal mengidentifikasi kelompok yang lebih rentan terhadap masalah jantung, dilakukan pemisahan antara label dan fitur terlebih dahulu. Kemudian, fitur dilatih menggunakan K-means dengan nilai *cluster* sebanyak 2, karena penulis ingin melihat performanya dalam membagi kelompok yang rentan dan tidak. Setelah itu, hasil *clustering* akan dibandingkan dengan label asli untuk melihat akurasi. Gambar 9 menunjukkan visualisasi hasil *clustering* menggunakan 2 *cluster*.

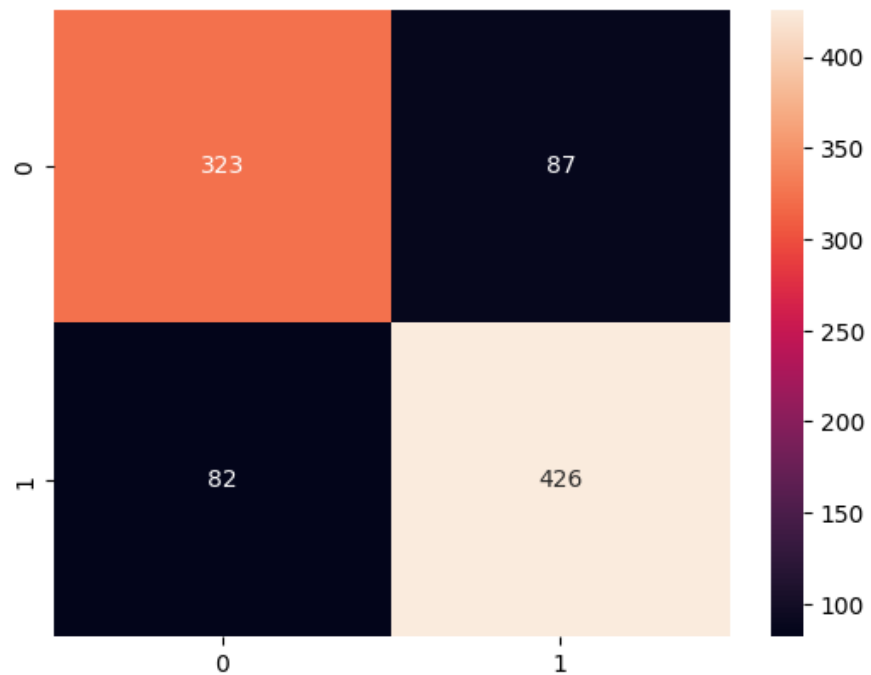
Gambar 9 Visualisasi K-Means dengan PCA



Karena fitur berdimensi tinggi, dilakukan reduksi dimensi terlebih dahulu menggunakan algoritma PCA. Hasilnya, seperti terlihat pada gambar 8, terdapat dua kelompok yang berbeda, bergantung pada *centroid* mana dia lebih dekat. Kemudian, dilakukan perbandingan antara label K-Means dengan label asli. Didapatkan nilai akurasi sebesar 18,4%, nilai yang sangat kecil. Akan tetapi, karena pada *clustering* model hanya membagi tanpa melihat labelnya, nilai akurasi yang kecil tersebut mungkin disebabkan karena label antara 0 dan 1 nya tertukar.

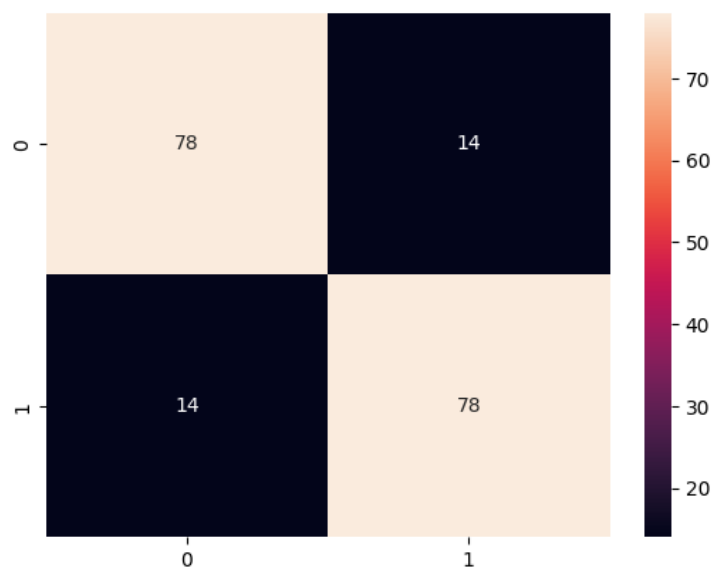
Setelah nilai label K-means ditukar, didapatkan akurasi sebesar 81,6%. Hasil ini terbilang cukup tinggi bagi algoritma *clustering* dalam membedakan dua kategori tertentu, yaitu pasien yang berpotensi mengalami masalah jantung dan yang tidak. Dengan demikian, dapat disimpulkan bahwa metode clustering, khususnya K-Means, berpotensi menjadi alat yang berguna dalam analisis dan pemantauan pasien terkait risiko masalah jantung. Gambar 10 menunjukkan *confusion matrix* K-Means

Gambar 10 *Confusion Matrix K-Means*



Setelah didapatkan hasil ini, penulis memutuskan untuk melakukan training kembali tetapi menggunakan data yang telah dipisahkan menjadi data training dan data validasi seperti pada metode sebelumnya. Hasilnya, akurasi training mencapai 80,8% dan akurasi validasi mencapai 84,8%. Hasil ini menunjukkan bahwa metode *clustering* bisa saja digunakan untuk klasifikasi dua kategori data. Hanya saja, terdapat kemungkinan label prediksi dapat terbalik. Gambar 11 menunjukkan *confusion matrix K-means* pada data validasi

Gambar 11 *Confusion matrix K-means* pada data validasi




d. Kesimpulan

Berdasarkan metode-metode yang telah dilakukan, terlihat bahwa metode Artificial Neural Network (ANN) menonjol sebagai metode yang memiliki performa paling baik. Model ANN berhasil mencapai akurasi sebesar 89.1% dalam memprediksi data validasi. Akurasi ini mencerminkan kemampuan model ANN dalam mengklasifikasikan pasien berdasarkan karakteristik yang terkandung dalam data medis. Hasil ini menunjukkan bahwa ANN dapat menjadi alat yang efektif dalam membantu dokter dan peneliti dalam pemantauan risiko masalah jantung pada pasien.

Namun, perlu dicatat bahwa pemilihan model tergantung pada berbagai faktor, termasuk tujuan analisis, sumber daya yang tersedia, dan ketersediaan data. Setiap metode, termasuk *Decision tree*, K-Means dan ANN, memiliki keunggulan dan kelemahan masing-masing. Oleh karena itu, pemilihan metode harus didasarkan pada kebutuhan dan konteks analisis yang spesifik.


Lampiran

Colab Decision tree

 Heart Failure Prediction Decision Tree.ipynb


https://colab.research.google.com/drive/1CEC8q-25RapEF_1amA1X7QsSH1KqIwOw?authuser=1

Colab ANN

 Heart Failure Prediction ANN.ipynb

https://colab.research.google.com/drive/1EOdl_zjeoAnjtckie-m2pf5GaNrEOiFP?usp=sharing

Colab K-Means

 Heart Failure Prediction K Means.ipynb

https://colab.research.google.com/drive/1Qk-wVRldkbPCx48XEIQ2VgWbYV_n_R6l?usp=sharing

Link Dataset

<https://drive.google.com/file/d/1eM4VSkt7JB67JJjNa6dA9wjbyCzIqFy4/view?usp=sharing>

Link Clean Dataset

<https://drive.google.com/file/d/12zDitOnAqSy47rZnAgJ-d6n3vaBepZFr/view?usp=sharing>