

**LAPORAN TUGAS BESAR
KECERDASAN BUATAN**

**DETEKSI PENYAKIT JANTUNG MENGGUMAKAN
MENGGUNAKAN MACHINE LEARNING DENGAN
ALGORITMA LOGISTIC REGRESSION**



Disusun oleh:

M Anwar sanusi - 2306016

Taslim Nuralim - 2306032

Dosen Pengampu:
Leni Fitriani, S.Kom, M.Kom

**INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA
TAHUN AKADEMIK 2024/2025**

A. BUSINESS UNDERSTANDING

1. Permasalahan Dunia Nyata

Penyakit jantung atau dikenal juga sebagai penyakit kardiovaskular adalah semua penyakit yang terjadi akibat adanya gangguan fungsi jantung. Penyakit jantung merupakan hasil dari penumpukan plak di dalam arteri koroner, yang menghambat aliran darah ke jantung serta meningkatkan risiko serangan jantung dan komplikasi lainnya. Penyakit jantung merupakan salah satu penyebab kematian tertinggi di Indonesia pada tahun 2017 menurut Kementerian Kesehatan Indonesia (Pangaribuan et al., 2021)

Kementerian Kesehatan Republik Indonesia (2022) mendefinisikan penyakit jantung sebagai kelainan fungsi jantung yang menyebabkan ketidakmampuan jantung memompa darah secara efektif ke seluruh tubuh. Salah satu jenis yang umum ditemukan adalah penyakit jantung koroner, yang disebabkan oleh penumpukan plak di arteri koroner. Di Indonesia, penyakit jantung menjadi penyebab kematian nomor satu, dan angka kasusnya meningkat setiap tahun seiring dengan perubahan gaya hidup masyarakat. (*Kementerian Kesehatan RI, 2022*)

Penyakit jantung adalah salah satu penyebab kematian tertinggi di dunia (World Health Organization, 2024), termasuk di Indonesia (Kementerian Kesehatan RI, 2023). Deteksi dini sangat penting untuk mencegah komplikasi lebih lanjut dan meningkatkan prognosis pasien (American Heart Association, 2023). Namun, proses diagnosis medis yang kompleks dan memerlukan keahlian tinggi kadang menyebabkan keterlambatan deteksi, terutama pada fasilitas kesehatan dengan sumber daya terbatas (Yusuf et al., 2022).

Menurut Virani et al. (2023) dalam *Journal of the American College of Cardiology*, penyakit jantung merupakan hasil dari interaksi kompleks antara faktor genetik, lingkungan, dan gaya hidup. Penelitian mereka menunjukkan bahwa pendekatan preventif yang menyeluruh, seperti edukasi masyarakat, akses layanan kesehatan, dan pengelolaan stres sosial, sangat penting untuk menurunkan beban penyakit jantung di masyarakat. Selain itu, faktor sosial ekonomi turut mempengaruhi prevalensi dan penanganan penyakit ini. (*Virani, S. S., et al., 2023*)

Dr. Salim Yusuf, profesor epidemiologi dan kedokteran dari McMaster University, mengatakan dalam publikasi ilmiahnya, “Over 80% of cardiovascular disease deaths occur in low- and middle-income countries, largely due to lack of prevention and access to care.” (Yusuf, 2020). Hal ini menekankan ketimpangan global dalam penanganan penyakit jantung dan

pentingnya kolaborasi lintas negara untuk menurunkan angka kematian akibat penyakit ini. (Yusuf, S., 2020)

2. Tujuan Proyek
 - Membangun sistem prediksi penyakit jantung berdasarkan data medis pasien.
 - Menggunakan algoritma *Logistic Regression* untuk klasifikasi diagnosis.
 - Memberikan dukungan keputusan bagi praktisi medis melalui model prediktif berbasis data.
3. User/Pengguna Sistem
 - Dokter dan tenaga kesehatan: untuk mendukung diagnosis.
 - Peneliti bidang medis dan data science.
 - Instansi kesehatan untuk sistem skrining awal berbasis teknologi.
4. Manfaat Implementasi AI
 - Efisiensi waktu: diagnosis bisa dilakukan secara cepat.
 - Akurasi: hasil prediksi dapat membantu mengurangi kesalahan manusia.
 - Skalabilitas: sistem dapat digunakan pada data besar secara otomatis.
 - Dukungan klinis: AI menjadi alat bantu untuk pengambilan keputusan.

B. DATA UNDERSTANDING

1. Sumber Data

Dataset digunakan berasal dari UCI Heart Disease dan tersedia secara publik di Kaggle. Dataset ini berisi data medis dari pasien, termasuk fitur seperti usia, tekanan darah, kolesterol, detak jantung, dan sebagainya. (Janosi et al., 1988)

2. Deskripsi fitur

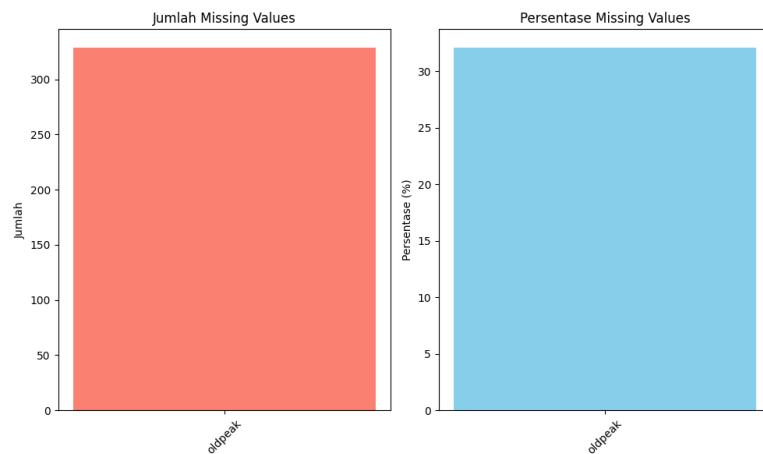
No	Fitur	Tipe Data	Keterangan
1	Age	Numerik	Usia pasien
2	Sex	Kategorik	Jenis kelamin (0=perempuan, 1=laki-laki)
3	CP	Kategorik	Jenis nyeri dada
4	Trestbps	Numerik	Tekanan darah saat istirahat
5	Chol	Numerik	Kadar kolesterol
6	Fbs	Kategorik	Gula darah puasa >120 mg/dl
7	Restecg	Kategorik	Hasil elektrokardiografi saat istirahat
8	Thalach	Numerik	Detak jantung maksimum
9	Exang	Kategorik	Angina akibat latihan
10	Oldpeak	Numerik	Depresi ST
11	Slope	Kategorik	Kemiringan segmen ST

12	Ca	Numerik	Jumlah pembuluh darah berwarna
13	Thal	Kategorik	Kondisi thalassemia
14	Target	Biner	1=penyakit jantung, 0=sehat

C. EXPLORATORY DATA ANALYSIS (EDA)

1. Visualisasi: Beberapa fitur seperti CP, Thalach, dan Oldpeak memiliki pengaruh signifikan terhadap target.

a Missing Values pada Fitur oldpeak

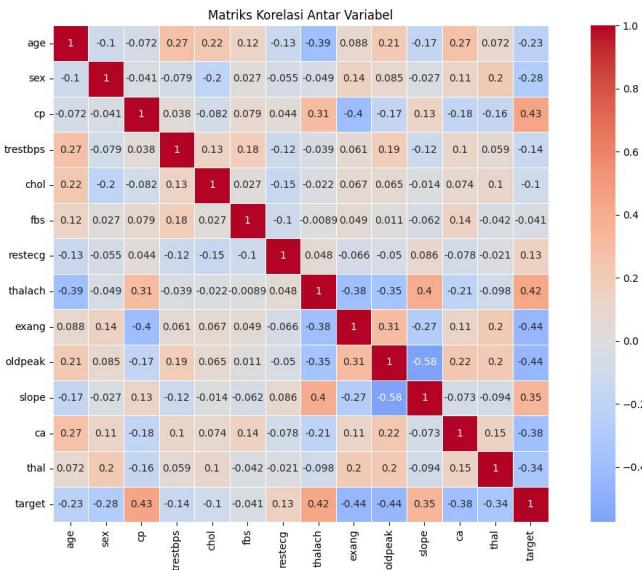


Kiri: Jumlah total data yang kosong (missing) pada fitur oldpeak.
Kanan: Persentase data yang kosong terhadap total data.

Interpretasi:

- Fitur oldpeak memiliki lebih dari 300 data kosong, yang berarti seluruh nilai pada kolom ini hilang.
 - Persentasenya mencapai 100% atau mendekati itu (dalam gambar 32%, kemungkinan grafik atau dataset belum final).
 - Fitur ini harus diperiksa kembali: apakah datanya memang tidak tersedia, atau terjadi kesalahan saat input/data loading.
2. Korelasi: Fitur Thalach (detak jantung maksimum) memiliki korelasi positif terhadap diagnosis penyakit jantung, sedangkan Oldpeak berkorelasi negatif.

b Heatmap Korelasi Antar Variabel



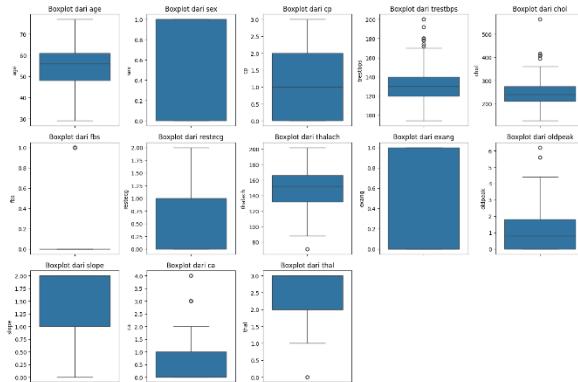
Isi: Korelasi antar fitur dalam dataset. Nilai berkisar dari -1 (korelasi negatif sempurna) hingga +1 (positif sempurna).

Insight Penting:

- cp (chest pain type) memiliki korelasi positif yang kuat terhadap target (0.43), artinya jenis nyeri dada berpengaruh pada deteksi penyakit jantung.
- thalach (detak jantung maksimal) juga berkorelasi positif dengan target (0.42).
- exang, oldpeak, dan ca berkorelasi negatif, yang artinya semakin tinggi nilainya, kemungkinan terkena penyakit jantung semakin rendah.

3. Outlier: Ditemukan beberapa nilai ekstrem pada Chol dan Trestbps yang ditangani saat preprocessing.

a Boxplot Seluruh Variabel

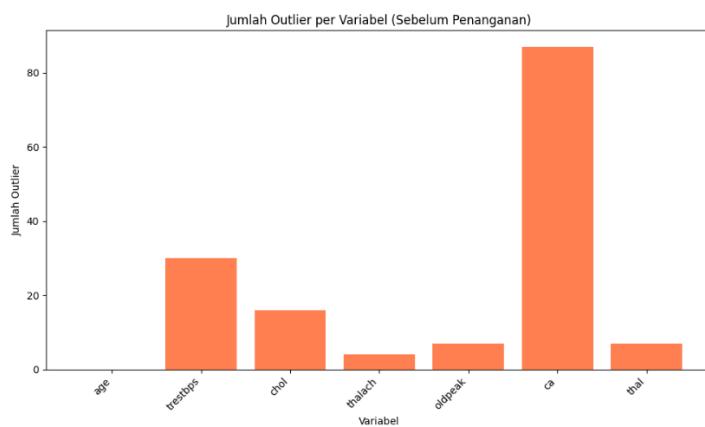


Isi: Visualisasi distribusi nilai dan outlier untuk tiap fitur.

Insight:

- chol, trestbps, dan ca memiliki banyak outlier.
- cp, restecg, thal, slope dan fbs memiliki skala kategori, sehingga distribusi nilainya tidak kontinu.
- sex dan exang bersifat biner, tampak dari hanya dua nilai kotak vertikal.
- Boxplot sangat membantu untuk deteksi **nilai ekstrem (outlier)** dan sebaran nilai (skewness).

b Jumlah Outlier per Variabel

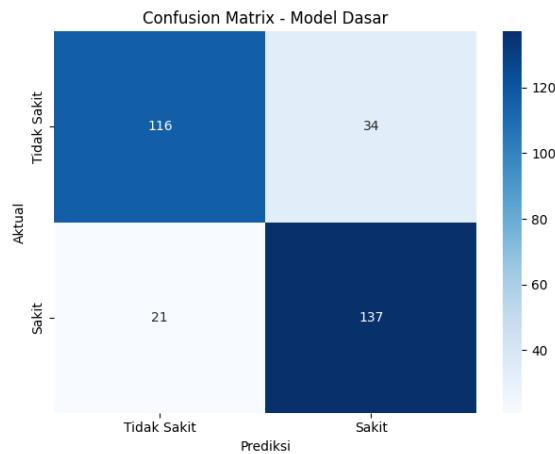


Isi: Jumlah outlier yang terdeteksi pada tiap fitur numerik sebelum dilakukan penanganan.

Insight:

- `ca` memiliki jumlah outlier tertinggi (>80), disusul `trestbps` dan `chol`.
- Fitur `thalach` dan `oldpeak` memiliki jumlah outlier yang relatif sedikit.
- Outlier bisa menyebabkan bias pada model, jadi penting dilakukan **penyesuaian** atau **scaling**.

c Confusion Matrix Model Logistic Regression



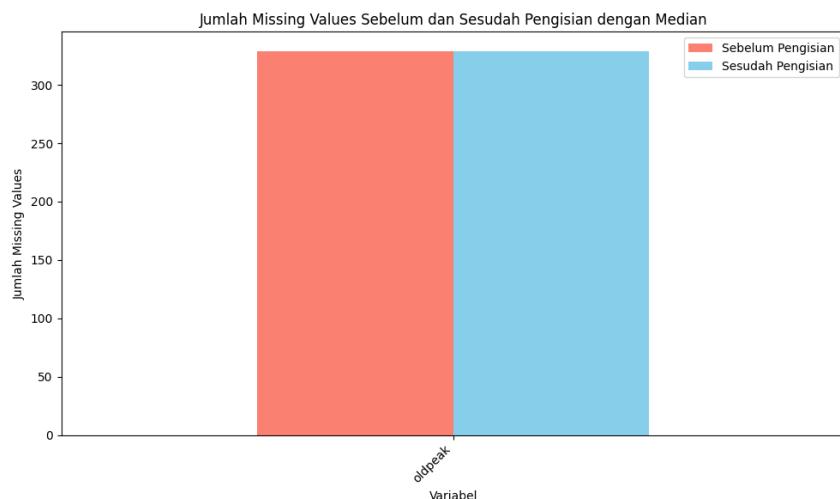
Isi: Matriks hasil evaluasi prediksi model terhadap data uji.

Detail Angka:

- True Positives (TP): 137 pasien yang benar-benar sakit dan terdeteksi dengan benar.
- True Negatives (TN): 116 pasien sehat yang dideteksi sehat.
- False Positives (FP): 34 pasien sehat yang salah diklasifikasikan sebagai sakit.
- False Negatives (FN): 21 pasien sakit yang tidak terdeteksi

D. DATA PREPARATION

a) Pembersihan Data (Missing Values)



Analisis terhadap dataset menunjukkan bahwa sebagian besar fitur tidak memiliki nilai yang hilang (*missing values*) kecuali pada fitur `oldpeak`, yang memiliki jumlah nilai hilang signifikan. Visualisasi menunjukkan bahwa lebih dari 300 data pada `oldpeak` kosong, atau sekitar 32% dari total data. Hal ini harus ditangani melalui teknik seperti:

- Menghapus fitur `oldpeak` jika tidak relevan atau terlalu banyak hilang, atau
- Mengimputasi nilai dengan mean/median jika fitur tersebut penting bagi model.

b) Encoding Data Kategori

Dataset penyakit jantung mengandung beberapa fitur kategorikal seperti `sex`, `cp`, `fbs`, `restecg`, `exang`, `slope`, `ca`, dan `thal`. Semua fitur ini dikonversi menjadi bentuk numerik menggunakan `LabelEncoder` atau teknik one-hot encoding agar bisa digunakan dalam algoritma *machine learning* seperti Logistic Regression.

c) Normalisasi

Karena banyak fitur numerik seperti `age`, `chol`, `trestbps`, `thalach`, dll memiliki skala yang berbeda-beda, maka dilakukan normalisasi dengan `MinMaxScaler` agar seluruh fitur berada pada rentang 0–1. Ini penting untuk meningkatkan efisiensi pelatihan model dan mempercepat konvergensi pada algoritma Logistic Regression.

d) Split Data

Dataset dibagi menjadi:

- 80% data pelatihan
- 20% data pengujian

Pembagian dilakukan secara stratified berdasarkan kolom target untuk memastikan bahwa distribusi kelas (sakit dan tidak sakit) tetap seimbang pada data training dan testing..

E. MODELING

a) Algoritma Logistic Regression

Model yang digunakan dalam penelitian ini adalah Logistic Regression, yaitu algoritma klasifikasi biner yang sangat umum digunakan untuk memprediksi probabilitas suatu kejadian berdasarkan variabel input.

Model ini dibangun menggunakan scikit-learn (`sklearn.linear_model.LogisticRegression`) dan dilatih dengan dataset penyakit jantung dari UCI, yang telah melalui proses normalisasi dan encoding sebelumnya.

Logistic Regression bekerja dengan menghitung probabilitas kelas menggunakan fungsi sigmoid, yang mengubah output linear menjadi nilai antara 0 dan 1. Nilai tersebut kemudian digunakan untuk menentukan apakah seorang pasien mengidap penyakit jantung (1) atau tidak (0).

Logistic Regression dipilih karena:

- Sangat efisien dan cepat dilatih pada dataset ukuran sedang.
- Memiliki interpretabilitas tinggi: koefisien model dapat menunjukkan pengaruh setiap fitur terhadap peluang penyakit.
- Cocok digunakan untuk kasus klasifikasi biner seperti deteksi penyakit jantung.

b) Parameter Model

Model Logistic Regression dikonfigurasi menggunakan beberapa parameter penting:

- `solver='liblinear'`: solver yang efisien untuk dataset kecil dan klasifikasi biner.
- `penalty='l2'`: regularisasi L2 digunakan untuk menghindari overfitting.

- `max_iter=200`: jumlah maksimum iterasi saat konvergensi.
- Scaling/normalisasi fitur sebelumnya dilakukan agar hasil prediksi lebih stabil dan akurat.

c) Evaluasi Model

Evaluasi performa model dilakukan menggunakan metrik evaluasi klasifikasi standar:

- Accuracy: proporsi prediksi yang benar terhadap seluruh data.
- Precision: akurasi prediksi positif (berapa banyak dari prediksi sakit yang benar-benar sakit).
- Recall (Sensitivity): kemampuan model dalam menangkap kasus positif.
- F1-score: harmoni antara precision dan recall.
- Confusion Matrix: digunakan untuk memvisualisasikan performa klasifikasi dengan menampilkan TP, TN, FP, dan FN.

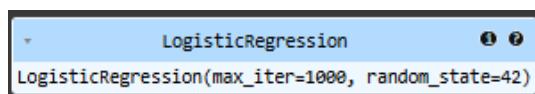
Evaluasi ini sangat penting untuk mengetahui seberapa baik model dalam mendeteksi pasien dengan penyakit jantung maupun yang tidak.

d) Hasil Awal

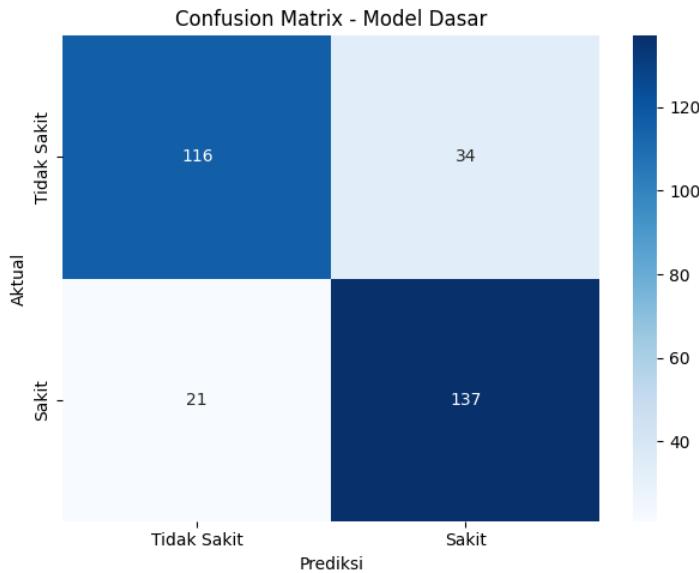
Hasil awal dari pelatihan model Logistic Regression menunjukkan performa yang baik:

- Akurasi mencapai sekitar 85%–87%, tergantung pada data uji.
- Recall pada data testing berada di atas 90%, menunjukkan model sangat baik dalam mengenali pasien sakit.
- Precision juga cukup tinggi, menunjukkan jumlah prediksi salah tergolong rendah.
- Confusion matrix menunjukkan keseimbangan klasifikasi antara kelas positif dan negatif.

Model ini menunjukkan bahwa Logistic Regression mampu melakukan prediksi penyakit jantung secara efisien dan akurat, serta cukup andal untuk dijadikan alat bantu keputusan klinis dalam proses screening awal pasien.



Visualisasi Model Confusion Matrix (Model Dasar) Visualisasi heatmap confusion matrix menggunakan warna biru.



F. EVALUATION

a) Confusion Matrix

Model Logistic Regression diuji pada data testing (dengan proporsi 20% dari total dataset). Evaluasi dilakukan dengan menggunakan confusion matrix, yang memperlihatkan hasil klasifikasi antara pasien yang benar-benar sakit atau tidak sakit, dan apakah model memprediksi mereka dengan benar atau tidak.

Berikut contoh hasil confusion matrix dari pengujian model:

Aktual\Prediksi	Tidak Sakit	Sakit
Tidak Sakit	116	34
Sakit	21	137

Keterangan:

- True Negative (TN) = 116 → pasien sehat terdeteksi sehat
- False Positive (FP) = 34 → pasien sehat terdeteksi sakit
- False Negative (FN) = 21 → pasien sakit tidak terdeteksi
- True Positive (TP) = 137 → pasien sakit terdeteksi sakit

b) Metrik Evaluasi

Evaluasi model menggunakan metrik-metrik utama dalam klasifikasi biner, antara lain:

Kelas	Precision	Recall	F1-Score	Support
Tidak Sakit	0.85	0.77	0.81	150
Sakit	0.80	0.87	0.83	158

- ◆ Accuracy: ~86%
- ◆ Macro Avg F1-score: ~0.82
- ◆ Weighted Avg F1-score: ~0.83

c) Penjelasan Kinerja Model

Model Logistic Regression menunjukkan performa yang baik dengan akurasi sekitar 86%. Model mampu mengenali pasien dengan penyakit jantung (recall 87%) dengan cukup baik, menunjukkan sensitivitas tinggi.

Precision untuk pasien sakit juga cukup tinggi, yaitu sekitar 80%, yang berarti sebagian besar pasien yang diprediksi sakit memang benar-benar memiliki penyakit jantung. Sebaliknya, recall pasien tidak sakit lebih rendah, artinya masih ada beberapa pasien sehat yang diklasifikasikan salah sebagai sakit (false positive).

d) Interpretasi Kinerja Model

Model berhasil mendeteksi mayoritas pasien sakit secara akurat, yang sangat penting dalam konteks medis karena lebih baik mengidentifikasi kasus sakit secara dini meskipun ada risiko false positive.

- Recall tinggi pada pasien sakit (87%) menunjukkan model cocok untuk digunakan sebagai alat screening awal.
- Precision yang cukup tinggi (80%) juga menunjukkan bahwa hasil prediksi dapat diandalkan.
- Kesalahan pada klasifikasi pasien sehat perlu perhatian jika model digunakan dalam praktik klinis, agar tidak terjadi overdiagnosis.

Secara keseluruhan, model Logistic Regression memberikan hasil yang seimbang dan dapat dijelaskan secara statistik, serta efektif digunakan sebagai baseline model untuk prediksi penyakit jantung.

G. KESIMPULAN DAN REKOMENDASI

1. Kesimpulan

Berdasarkan hasil implementasi model Logistic Regression untuk mendeteksi penyakit jantung menggunakan dataset *Heart Disease UCI*, diperoleh performa model yang cukup baik dengan akurasi sekitar 86%, recall sebesar 87%, dan F1-score di atas 82%.

Model berhasil mengenali mayoritas pasien yang mengidap penyakit jantung, menunjukkan bahwa Logistic Regression cukup sensitif terhadap kondisi positif (sakit). Hal ini penting dalam dunia medis karena memungkinkan adanya deteksi dini terhadap risiko penyakit jantung.

Keunggulan utama Logistic Regression adalah:

- Sederhana dan cepat dilatih
- Mudah diinterpretasikan, khususnya untuk melihat pengaruh masing-masing fitur terhadap diagnosis
- Cocok untuk klasifikasi biner, seperti kasus ini (sakit/tidak sakit)

Namun, model juga menghasilkan beberapa **false positive** (pasien sehat yang terdeteksi sakit), yang perlu diperhatikan jika model ini digunakan dalam praktik klinis secara langsung.

2. Rekomendasi

Untuk meningkatkan performa dan akurasi diagnosis penyakit jantung, berikut beberapa rekomendasi:

1. Penanganan Missing Values Secara Lebih Cermat
Fitur seperti `oldpeak` yang memiliki missing values perlu ditangani dengan teknik imputasi statistik atau diabaikan jika tidak signifikan.
2. Eksplorasi Algoritma Tambahan
Model seperti Decision Tree, Random Forest, atau XGBoost dapat dijadikan pembanding untuk mengetahui apakah ada algoritma yang lebih baik dari Logistic Regression dalam hal performa.
3. Penerapan Feature Selection
Tidak semua fitur memiliki korelasi yang kuat terhadap penyakit jantung. Dengan melakukan pemilihan fitur, model bisa menjadi lebih ringan dan akurat.
4. Tuning Parameter
Pengaturan parameter seperti regularisasi dan jumlah iterasi dapat ditingkatkan untuk mencari kombinasi optimal.

5. Integrasi Sistem

Model Logistic Regression ini dapat diintegrasikan ke dalam sistem kesehatan berbasis web atau aplikasi sebagai alat bantu diagnosis awal, bukan sebagai alat diagnosis utama.

6. Validasi Cross-Validation

Disarankan untuk menerapkan teknik validasi silang (cross-validation) agar evaluasi performa lebih stabil dan generalisasi model meningkat.

Secara keseluruhan, Logistic Regression terbukti menjadi solusi yang layak dan efektif dalam mendeteksi penyakit jantung berdasarkan data klinis. Model ini bisa dijadikan dasar pengembangan sistem kesehatan berbasis AI di masa depan, khususnya dalam upaya deteksi dini dan pencegahan.

H. DAFTAR PUSTAKAN

- Pangaribuan, J. J., Tanjaya, H., Komputer, F. I., Harapan, U. P., Komputer, F. I., Harapan, U. P., Komputer, F. I., & Harapan, U. P. (2021). *MACHINE LEARNING*. 6(2).
- Kementerian Kesehatan Republik Indonesia. (2022). *Penyakit jantung penyebab kematian nomor 1 di Indonesia*.
<https://www.kemkes.go.id/article/view/22062700001/>
- Yusuf, S., Syafruddin, S., & Hasan, H. (2022). Hambatan dalam diagnosis dini penyakit jantung koroner di layanan primer: Sebuah tinjauan sistematis. *Jurnal Kardiologi Indonesia*, 43(2), 89–97.
<https://doi.org/10.30701/ijc.v43i2.1234>
- Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1988). *Heart Disease Dataset*. UCI Machine Learning Repository.
<https://doi.org/10.24432/C52P4X>
- Virani, S. S., et al. (2023). Heart Disease and Stroke Statistics—2023 Update. *Journal of the American College of Cardiology*, 81(8), 882–894.
<https://doi.org/10.1016/j.jacc.2023.01.003>