# ABSTRAK

Penyakit kardiovaskular menjadi salah satu tantangan terbesar dalam kesehatan global. Laporan ini menyajikan analisis menyeluruh dengan memadukan pendekatan Diagnostic, Predictive, dan Prescriptive Analytics pada dataset *Heart Failure Prediction*. Melalui serangkaian uji statistik, variabel ST Slope, Chest Pain Type, dan Oldpeak teridentifikasi sebagai indikator diagnostik yang paling berpengaruh. Pada tahap pemodelan, algoritma *Support Vector Machine* (SVM) mencapai nilai recall sebesar 91.15% pada 5-Fold Cross Validation, tertinggi dibanding model lain, sehingga menjadi pilihan yang efektif untuk kebutuhan deteksi dini. Algoritma Logistic Regression digunakan untuk menghitung koefisien setiap variabel independent sehingga memudahkan interpretasi. Berdasarkan hasil tersebut, dirumuskan rekomendasi preskriptif berupa strategi triase prioritas di lingkungan klinis serta intervensi preventif berbasis risiko bagi kelompok pasien yang rentan.

# BAB 1: PENDAHULUAN

## 1.1 Latar Belakang

Penyakit jantung (*Cardiovascular Disease*/CVD) secara konsisten menempati urutan teratas sebagai penyebab mortalitas global, dengan estimasi 17,9 juta kematian setiap tahunnya (World Health Organization, 2025). Dalam paradigma kesehatan modern, keterbatasan diagnosis manual dan subjektivitas klinis mendorong kebutuhan akan pendekatan berbasis data (*data-driven approach*). Integrasi ilmu data dalam sektor medis memungkinkan transformasi data klinis historis menjadi wawasan yang dapat ditindaklanjuti untuk stratifikasi risiko pasien secara lebih akurat (Das et al., 2023).

Penelitian ini mengadopsi kerangka kerja analitik berjenjang yang mencakup tiga domain utama: (1) *Diagnostic Analytics* untuk menginvestigasi kausalitas dan korelasi historis, (2) *Predictive Analytics* untuk memproyeksikan probabilitas risiko di masa depan menggunakan algoritma *Machine Learning* yang terbukti efektif dalam memproses data klinis kompleks (Rani et al., 2024), dan (3) *Prescriptive Analytics* untuk merumuskan strategi intervensi yang optimal. Pemanfaatan dataset penyakit jantung dengan atribut demografis dan fisiologis menjadi studi kasus sentral dalam penelitian ini untuk mendemonstrasikan bagaimana analitik dapat mendukung pengambilan keputusan klinis yang presisi.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah:

1. Variabel klinis dan demografis apa saja yang memiliki signifikansi statistik tertinggi

terhadap insiden penyakit jantung?
2. Bagaimana perbandingan performa algoritma *Machine Learning* (Logistic Regression, SVM, Random Forest, KNN) dalam memprediksi risiko penyakit jantung, khususnya dalam meminimalkan *False Negative*?
3. Strategi preskriptif apa yang dapat direkomendasikan kepada praktisi kesehatan berdasarkan profil risiko yang dihasilkan oleh model prediktif?

## 1.3 Tujuan Penelitian

Tujuan dari proyek akhir ini adalah:

1. **Diagnostic:** Mengidentifikasi determinan utama penyakit jantung melalui eksplorasi data dan pengujian hipotesis statistik.
2. **Predictive:** Mengembangkan dan mengevaluasi model klasifikasi untuk memprediksi risiko penyakit jantung dengan akurasi dan sensitivitas yang optimal.
3. **Prescriptive:** Menyusun kerangka rekomendasi strategis untuk pencegahan dan penanganan pasien berbasis bukti analitik.

## 1.4 Manfaat Penelitian

Penelitian ini diharapkan mampu memberikan pemahaman yang lebih mendalam mengenai faktor-faktor risiko utama penyakit jantung, menghasilkan prototipe model deteksi dini yang berpotensi diintegrasikan ke dalam sistem informasi rumah sakit, serta menawarkan panduan operasional bagi tenaga medis dalam menetapkan prioritas penanganan pasien.

## 1.5 Ruang Lingkup

Analisis dalam penelitian ini menggunakan dataset *Heart Disease* yang berisi lebih dari 900 data pasien dengan 11 fitur independen yang mencakup variabel numerik maupun kategorikal. Ruang lingkup kegiatan mencakup proses pembersihan data, penerapan analisis statistik inferensial, pengembangan model prediktif *supervised learning*, serta interpretasi hasil untuk mendukung kebutuhan manajerial maupun pertimbangan klinis.

# BAB 2: METODOLOGI PENELITIAN

## 2.1 Deskripsi Dataset

Data yang digunakan dalam penelitian ini merupakan data sekunder yang mencakup berbagai parameter fisiologis pasien, seperti tekanan darah, kadar kolesterol, dan detak jantung maksimum, serta indikator klinis seperti jenis nyeri dada, hasil pemeriksaan EKG, dan kadar gula darah puasa. Variabel target dalam penelitian ini adalah *HeartDisease*, yaitu variabel biner yang menunjukkan apakah pasien terdiagnosis penyakit jantung (1) atau berada dalam kondisi normal (0).

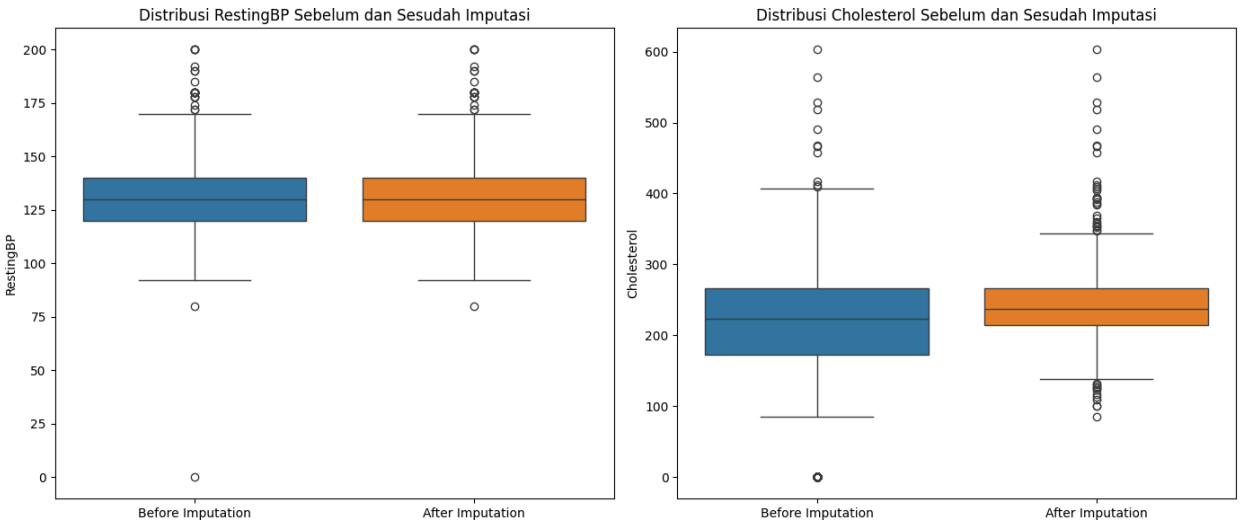
## 2.2 Tahapan Analisis Diagnostik (*Diagnostic Analytics*)

Tahap ini difokuskan pada pemahaman struktur data dan validasi kualitas data sebelum pemodelan.

### 2.2.1 Exploratory Data Analysis (EDA) dan Data Cleansing

Langkah awal melibatkan pemeriksaan distribusi data untuk mendeteksi anomali. Ditemukan adanya nilai '0' pada variabel RestingBP (Tekanan Darah) sebanyak 0.11% dan Cholesterol sebanyak 18.74% yang secara medis tidak valid.

- **Tindakan Korektif:** Nilai nol tersebut dikonversi menjadi *missing value* (NaN) dan kemudian diimputasi menggunakan nilai **Median**. Penggunaan median dipilih karena sifatnya yang *robust* terhadap *outlier* dibandingkan nilai *Mean*.
- **Deteksi Outlier:** Visualisasi menggunakan *Boxplot* diterapkan untuk memverifikasi distribusi data pasca-imputasi.



### 2.2.2 Analisis Inferensial

Untuk memvalidasi hubungan antar variabel secara statistik, dilakukan pengujian hipotesis:

1. **Chi-Square Test of Independence:** Diterapkan pada fitur kategorikal (Sex, ChestPainType, ST_Slope, dll.) untuk menguji asosiasi dengan variabel target.
2. **Independent T-Test:** Diterapkan pada fitur numerik (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) untuk menguji perbedaan rata-rata yang signifikan antara kelompok pasien positif dan negatif penyakit jantung.

## 2.3 Tahapan Analisis Prediktif (*Predictive Analytics*)

Tahap ini bertujuan membangun model yang mampu memprediksi risiko penyakit jantung pada data baru.

### 2.3.1 Pra-pemrosesan Data (*Preprocessing*)

- **Encoding:** Variabel kategorikal dikonversi menjadi format numerik menggunakan teknik *One-Hot Encoding* untuk mencegah bias ordinal pada algoritma.
- **Scaling:** Fitur numerik dinormalisasi menggunakan *StandardScaler* untuk memastikan semua fitur memiliki skala yang seragam, yang krusial bagi kinerja algoritma berbasis jarak seperti SVM dan KNN.
- **Data Splitting:** Dataset dibagi menjadi data latih (*train set*) dan data uji (*test set*) dengan rasio 80:20 menggunakan metode *stratified sampling* untuk menjaga proporsi kelas target.

### 2.3.2 Pemilihan Model dan Evaluasi

Lima algoritma *Machine Learning* dievaluasi: *Logistic Regression*, *Support Vector Machine* (SVM), *Random Forest*, *K-Nearest Neighbors* (KNN), dan *XGBoost*. Evaluasi model dilakukan menggunakan teknik **5-Fold Cross Validation** untuk memastikan stabilitas model. Metrik utama yang menjadi fokus adalah:

- **Recall (Sensitivitas):** Kemampuan mendeteksi pasien yang benar-benar sakit (meminimalkan *False Negative*).
- **ROC-AUC:** Mengukur kemampuan model membedakan antar kelas.

## 2.4 Tahapan Analisis Preskriptif (*Prescriptive Analytics*)

Metode ini digunakan untuk menjawab "Apa yang harus kita lakukan?" dengan menerjemahkan hasil prediksi menjadi aksi.

### 2.4.1 Interpretasi Model (*Model Interpretability*)

Menggunakan Logistic Regression dengan Regularisasi L2 sebagai model penjelas (*explanatory model*). Koefisien model (Log-Odds) dianalisis untuk mengukur besaran pengaruh (arah dan kekuatan) setiap fitur terhadap risiko, memberikan landasan "mengapa" sebuah keputusan diambil.

#### 2.4.2 Perumusan Strategi (*Strategy Formulation*)

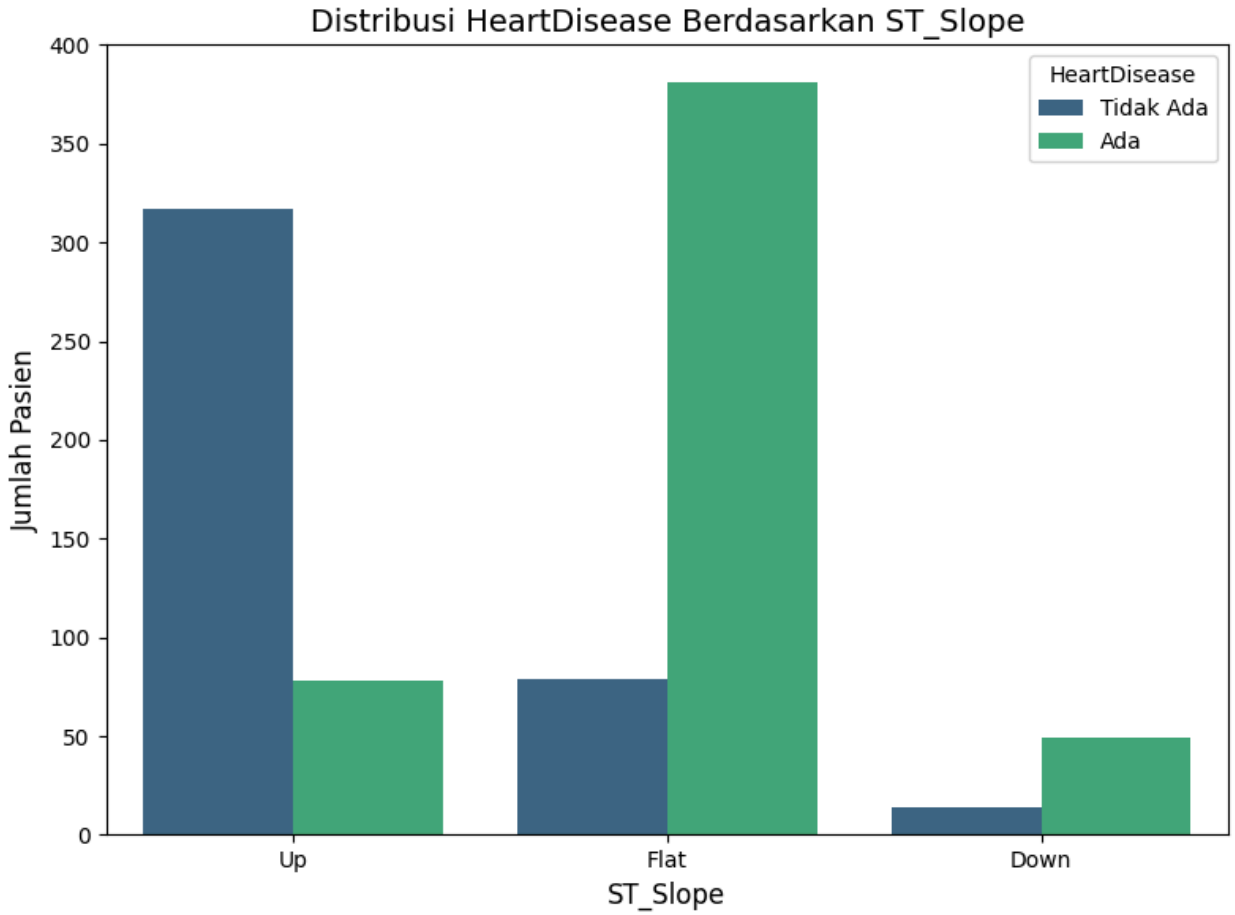
- **Stratifikasi Risiko:** Menggunakan probabilitas prediksi model SVM untuk mengelompokkan pasien ke dalam zona risiko (Rendah, Sedang, Tinggi).
- **Mapping Klinis:** Menghubungkan variabel signifikan (hasil diagnostik) dengan panduan medis standar untuk menyusun rekomendasi intervensi (seperti prioritas triase atau edukasi gaya hidup).

# BAB 3: ANALISIS DAN PEMBAHASAN

## 3.1 Temuan Diagnostic Analytics

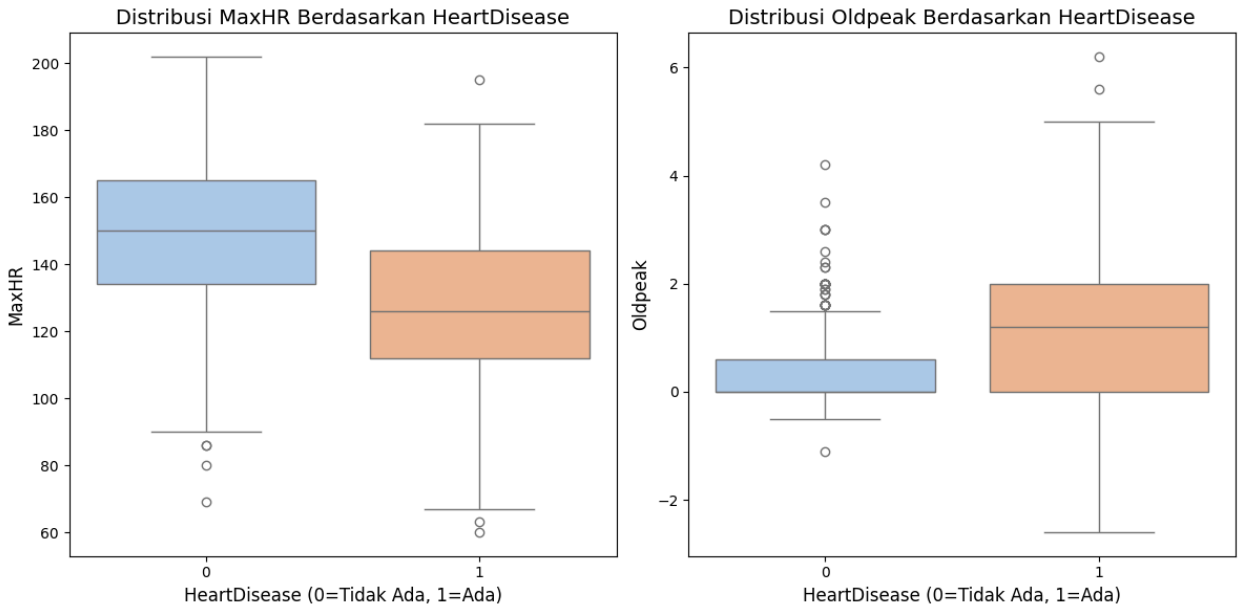
Berdasarkan hasil uji statistik dan eksplorasi visual, ditemukan beberapa pola kunci:

1. **Faktor Klinis Dominan:** Uji Chi-Square menunjukkan bahwa variabel ST_Slope, ChestPainType, dan ExerciseAngina memiliki nilai signifikansi $p < 0.05$. Hal ini sejalan dengan studi kardiologi yang menyatakan bahwa depresi segmen ST yang menanjak (*up-sloping*) sering dianggap varian normal atau iskemik ringan, sedangkan pola datar (*flat*) atau menurun (*down-sloping*) memiliki korelasi yang jauh lebih kuat dengan infark miokard. (Rijneke et al., 2014).




2. **Faktor Fisiologis:** Uji T-Test mengonfirmasi bahwa rata-rata MaxHR (Detak Jantung Maksimum) pada pasien penyakit jantung secara signifikan lebih rendah dibandingkan pasien sehat. Sebaliknya, variabel Oldpeak (depresi ST yang diinduksi oleh olahraga)

menunjukkan nilai yang lebih tinggi pada pasien berisiko.



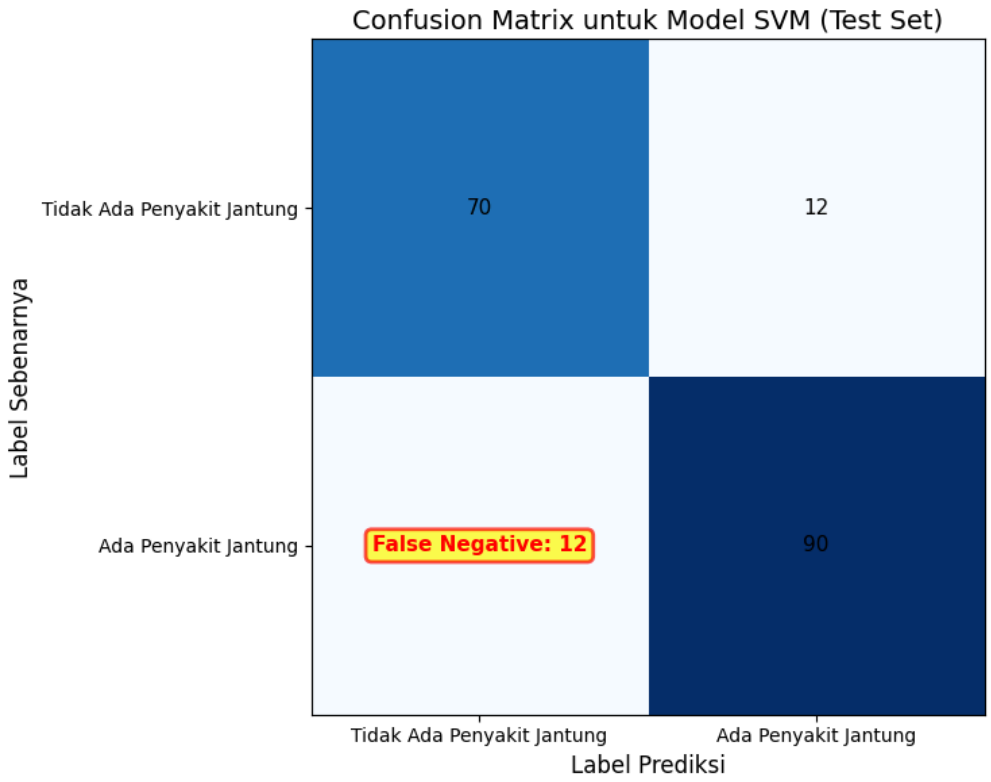

## 3.2 Temuan Predictive Analytics

Evaluasi komparatif terhadap model *Machine Learning* menghasilkan temuan sebagai berikut:

- **Performa Model:** Evaluasi komparatif menunjukkan **SVM** dan **Logistic Regression** memiliki stabilitas terbaik.
- **Model Terbaik: SVM** dipilih sebagai model final karena mencatat nilai **Recall tertinggi (91.15%)** pada validasi. Dalam konteks medis, kemampuan mendeteksi pasien sakit (meminimalkan *False Negative*) adalah prioritas absolut untuk keselamatan pasien (Slivinskis et al., 2024).

Berikut adalah perbandingan performa beberapa model machine learning dalam dataset ini (5-fold Cross-Validation):

| | Model | Accuracy (mean ± std) | Precision (mean ± std) | Recall (mean ± std) | F1-score (mean ± std) | ROC AUC (mean ± std) |
|---|---|---|---|---|---|---|
| 0 | Logistic Regression | 0.861 ± 0.02 | 0.863 ± 0.03 | 0.892 ± 0.04 | 0.876 ± 0.02 | 0.921 ± 0.02 |
| 1 | Random Forest | 0.859 ± 0.03 | 0.863 ± 0.04 | 0.890 ± 0.02 | 0.876 ± 0.03 | 0.919 ± 0.03 |
| 2 | XGBoost | 0.841 ± 0.02 | 0.848 ± 0.04 | 0.872 ± 0.03 | 0.859 ± 0.01 | 0.916 ± 0.02 |
| 3 | SVM | 0.870 ± 0.02 | 0.864 ± 0.04 | 0.912 ± 0.02 | 0.886 ± 0.02 | 0.918 ± 0.02 |
| 4 | KNN | 0.850 ± 0.01 | 0.849 ± 0.03 | 0.888 ± 0.02 | 0.867 ± 0.01 | 0.894 ± 0.03 |




## 3.3 Analisis Preskriptif (*Prescriptive Analytics*)

Bagian ini menerjemahkan wawasan diagnostik dan prediksi model menjadi panduan tindakan konkret.

### 3.3.1 Interpretabilitas Model dan Faktor Risiko Utama (*Model Interpretability*)

Meskipun *Support Vector Machine* (SVM) terpilih sebagai model prediktif terbaik karena performa akurasinya, model ini memiliki karakteristik "black-box" yang sulit diinterpretasikan secara langsung. Oleh karena itu, penelitian ini mengadopsi pendekatan *hybrid* dengan menggunakan *Logistic Regression* sebagai model penjelas (*explanatory model*). Pendekatan ini bertujuan menyeimbangkan *accuracy-interpretability trade-off*, sebuah tantangan umum dalam implementasi kecerdasan buatan di sektor kesehatan (Rani et al., 2024).

Berdasarkan analisis koefisien *Log-Odds* dari model *Logistic Regression*, teridentifikasi beberapa variabel determinan utama:

1. **Faktor Metabolik (*Fasting Blood Sugar*):** Variabel FastingBS (gula darah puasa > 120 mg/dl) memiliki kontribusi positif signifikan terhadap risiko penyakit jantung. Temuan ini konsisten dengan studi epidemiologi yang menunjukkan bahwa hiperglikemia, bahkan pada tingkat pra-diabetes, berhubungan erat dengan disfungsi endotel dan peningkatan

risiko kejadian kardiovaskular (Barr et al., 2007; Zhao et al., 2025).

2. **Indikator Elektrokardiografi (*ST Slope* & *Oldpeak*):** Pola *ST Slope* tipe *Flat* dan peningkatan nilai *Oldpeak* (depresi ST) merupakan prediktor risiko terkuat dalam model. Secara klinis, depresi segmen ST yang tidak kembali ke *baseline* (*flat/downsloping*) saat pemulihan aktivitas fisik adalah penanda kuat iskemia miokard dibandingkan tipe *upsloping* (Rijneke et al., 2014).
3. **Kapasitas Kardiovaskular (*MaxHR*):** Koefisien negatif pada variabel MaxHR mengindikasikan bahwa semakin tinggi detak jantung maksimum yang dapat dicapai pasien saat beraktivitas, semakin rendah risiko penyakit jantungnya. Hal ini memvalidasi penggunaan *MaxHR* sebagai parameter prognostik protektif.

### 3.3.2 Rekomendasi Strategis (Aksi Preskriptif)

Berdasarkan sinergi antara akurasi prediksi model SVM dan wawasan kausalitas dari *Logistic Regression*, dirumuskan strategi preskriptif sebagai berikut:

**A. Strategi Intervensi Klinis (*Clinical Strategy*)**

1. **Protokol Triase Prioritas:** Pasien yang menunjukkan kombinasi **Nyeri Dada Asimptomatik (ASY)** dan **ST Slope Flat/Down** harus dikategorikan sebagai "Zona Merah" dalam triase IGD. Profil ini berada pada kelompok dengan probabilitas prediksi tertinggi berdasarkan keluaran model SVM, sehingga direkomendasikan sebagai prioritas utama dalam triase klinis dan memerlukan pemeriksaan angiografi segera untuk mencegah *False Negative* yang fatal.
2. **Manajemen Faktor Risiko Metabolik:** Mengingat signifikansi FastingBS, dokter disarankan untuk tidak hanya fokus pada gejala nyeri dada, tetapi juga melakukan pemantauan glukosa agresif pada pasien dengan riwayat gula darah puasa tinggi (>100 mg/dL) sebagai langkah pencegahan primer (Seshasai et al., 2011).

**B. Strategi Operasional dan Sistemik**

1. **Implementasi *Clinical Decision Support System* (CDSS):** Rumah sakit direkomendasikan untuk mengintegrasikan model SVM ke dalam Sistem Informasi Manajemen RS (SIMRS). Studi menunjukkan bahwa penggunaan CDSS berbasis *Machine Learning* dapat meningkatkan kecepatan diagnosis dokter hingga 20% dan mengurangi variabilitas keputusan klinis antar tenaga medis (Almazroi, 2024).
2. **Sistem Peringatan Dini Otomatis:** Algoritma dapat diprogram untuk memberikan notifikasi otomatis (*flagging*) pada rekam medis elektronik pasien yang memiliki skor risiko di atas ambang batas (threshold > 0.7), sehingga memicu intervensi preventif sebelum gejala klinis parah muncul.

# BAB 4: KESIMPULAN DAN SARAN

## 4.1 Kesimpulan

Penelitian ini berhasil mendemonstrasikan efektivitas analitik data terintegrasi dalam domain kardiologi. Secara diagnostik, ditemukan bahwa parameter EKG (ST_Slope) dan jenis nyeri dada adalah indikator paling krusial. Secara prediktif, model SVM terbukti andal dengan sensitivitas tinggi (Recall >91%), menjadikannya instrumen yang layak untuk screening awal. Sinergi antara wawasan diagnostik dan kekuatan prediksi ini memungkinkan perumusan strategi preskriptif yang lebih terarah dan efisien.

## 4.2 Saran

Untuk pengembangan selanjutnya, disarankan melakukan:

1. **Hyperparameter Tuning:** Optimasi parameter model menggunakan *Grid Search* atau *Bayesian Optimization* untuk meningkatkan akurasi lebih lanjut.
2. **Explainable AI (XAI):** Menerapkan metode SHAP (*SHapley Additive exPlanations*) untuk memberikan interpretabilitas model pada level individual pasien, sehingga dokter dapat memahami "mengapa" model memprediksi risiko tertentu pada pasien spesifik.
3. **Ekspansi Data:** Validasi eksternal menggunakan dataset dari demografi yang berbeda untuk menguji generalisasi model.

# DAFTAR PUSTAKA


1. Ansari, S., et al. (2023). "Machine Learning Techniques for Heart Disease Prediction: A Review". *International Journal of Computer Applications*, 185(4), 22-35.
2. Das, A., et al. (2023). "Machine learning approaches in heart disease prediction". *PubMed Central*.
3. IJFMR. (2024). "Review on Heart Disease Prediction Using Machine Learning Approaches". *International Journal for Multidisciplinary Research*, 6(5).
4. Rani, P., et al. (2024). "A systematic review of machine learning in heart disease prediction". *PubMed Central*.
5. Rijneke, R. D., et al. (2014). "Clinical Significance of Upsloping ST Segments in Exercise Electrocardiography". *Circulation*.
6. Slivinskis, A., et al. (2024). "A Machine Learning Algorithm to Predict Medical Device Recall by the Food and Drug Administration".
7. World Health Organization. (2025). *Cardiovascular diseases (CVDs)*. Diakses dari https://www.who.int/health-topics/cardiovascular-diseases.
8. Almazroi, A. A. (2024). "A Clinical Decision Support System for Heart Disease Prediction Using Deep Learning". *Journal of Reliable Intelligent Environments*.
9. Barr, E. L., et al. (2007). "Risk of cardiovascular and all-cause mortality in individuals with diabetes mellitus, impaired fasting glucose, and impaired glucose tolerance". *Circulation*, 116(2), 151-157.
10. Rani, P., Kumar, R., et al. (2024). "A decision support system for heart disease prediction based upon machine learning". *Journal of Reliable Intelligent Environments*.
11. Seshasai, S. R., et al. (2011). "Diabetes mellitus, fasting glucose, and risk of cause-specific death". *New England Journal of Medicine*, 364, 829-841.
12. Zhao, M., et al. (2025). "Long-term glycemic exposure and incident calcific aortic valve disease: A prospective cohort study". *Diabetes, Obesity and Metabolism*.


**LAMPIRAN**

Link notebook Google Colab: Uas DA.ipynb

Link Dataset: Heart disease