

**PERBANDINGAN PERFORMA MODEL RANDOM  
FOREST, XGBOOST, DAN LOGISTIC REGRESSION DALAM  
MENGANALISIS PENYAKIT JANTUNG**

**SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**BRILYAN HARWAN**

**22.11.5177**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2026**

**PERBANDINGAN PERFORMA MODEL RANDOM FOREST,  
XGBOOST, DAN LOGISTIC REGRESSION DALAM  
MENGANALISIS PENYAKIT JANTUNG**

**SKRIPSI**

untuk memenuhi salah satu syarat mencapai derajat Sarjana  
Program Studi Informatika



disusun oleh

**BRILYAN HARWAN**

**22.11.5177**

Kepada

**FAKULTAS ILMU KOMPUTER  
UNIVERSITAS AMIKOM YOGYAKARTA  
YOGYAKARTA**

**2026**

**HALAMAN PERSETUJUAN**

**SKRIPSI**

**PERBANDINGAN PERFORMA MODEL RANDOM FOREST,  
XGBOOST, DAN LOGISTIC REGRESSION DALAM  
MENGANALISIS PENYAKIT JANTUNG**

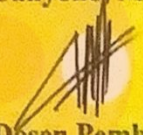
yang disusun dan diajukan oleh

**Brilyan Harwan**

**22.11.5177**

telah disetujui oleh Dosen Pembimbing Skripsi  
pada tanggal 23 Februari 2026

**Nuri Cahyono, M.Kom.**

  
**Nama Dosen Pembimbing**  
**NIK. 190302278**

**HALAMAN PENGESAHAN**  
**SKRIPSI**  
**PERBANDINGAN PERFORMA MODEL RANDOM FOREST,**  
**XGBOOST, DAN LOGISTIC REGRESSION DALAM**  
**MENGANALISIS PENYAKIT JANTUNG**

yang disusun dan diajukan oleh

**Brilyan Harwan**

**22.11.5177**

Telah dipertahankan di depan Dewan Penguji  
pada tanggal 23 Februari 2026

**Susunan Dewan Penguji**

**Nama Penguji**

**Tanda Tangan**

**Dr. Emigawaty, M.Kom**  
**NIK. 190302226**

**Ria Andriani, S.Kom., M.Kom**  
**NIK. 190302458**

**Nuri Cahyono, M.Kom**  
**NIK. 190302278**

Skripsi ini telah diterima sebagai salah satu persyaratan  
untuk memperoleh gelar Sarjana Komputer  
Tanggal 23 Februari 2026

**DEKAN FAKULTAS ILMU KOMPUTER**



**Prof. Dr. Kusrini, M.Kom.**  
**NIK. 190302106**

## HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Brilyan Harwan  
NIM : 22.11.5177

Menyatakan bahwa Skripsi dengan judul berikut:

### **Perbandingan Performa Model Random Forest, XGBoost, Logistic Regression dalam Menganalisis Penyakit Jantung**

Dosen Pembimbing : Nuri Cahyono, M.Kom.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 23 Februari 2026

Yang Menyatakan,



Brilyan Harwan

## HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur dan penyerahan, skripsi ini saya persembahkan kepada:

1. Tuhan Yesus Kristus yang telah memberikan saya kekuatan, kesabaran yang melimpah, serta kesehatan tubuh dan jiwa sehingga saya dapat melewati masa-masa sulit dalam proses penyusunan skripsi ini. Segala kemuliaan hanya bagi-Nya.
2. Papa dan mama serta seluruh keluarga yang selalu mendoakan, mendukung, dan menjadi pilar kekuatan di balik setiap usaha yang telah saya lakukan hingga skripsi ini selesai.
3. Diri saya sendiri yang telah berusaha, bangkit dari setiap keterpurukan, melawan rasa malas dan terus melangkah meski terkadang terasa berat.
4. Teman-teman dan sahabat saya di Yogyakarta yang telah menjadi keluarga kedua saya, yang selalu memberikan dukungan, tawa bersama, diskusi hingga larut malam, serta semangat yang tak pernah pudar sepanjang perjalanan ini.

## KATA PENGANTAR

Bagian ini berisi pernyataan resmi yang ingin disampaikan oleh penulis kepada pihak lain, misalnya ucapan terima kasih kepada Dosen Pembimbing, Tim Dosen Penguji, dan semua pihak yang terkait dalam penyelesaian skripsi termasuk orang tua dan penyandang dana.

Nama harus ditulis secara lengkap termasuk gelar akademik dan harus dihindari ucapan terima kasih kepada pihak yang tidak terkait. Bahasa yang digunakan harus mengikuti kaidah bahasa Indonesia yang baku.

Bagian ini tidak perlu dituliskan hal-hal yang bersifat ilmiah. Kata Pengantar diakhiri dengan mencantumkan kota dan tanggal penulisan diikuti di bawahnya dengan **kata “Penulis” tanpa perlu menyebutkan nama dan tanda tangan.**

Yogyakarta, 23 Februari 2026

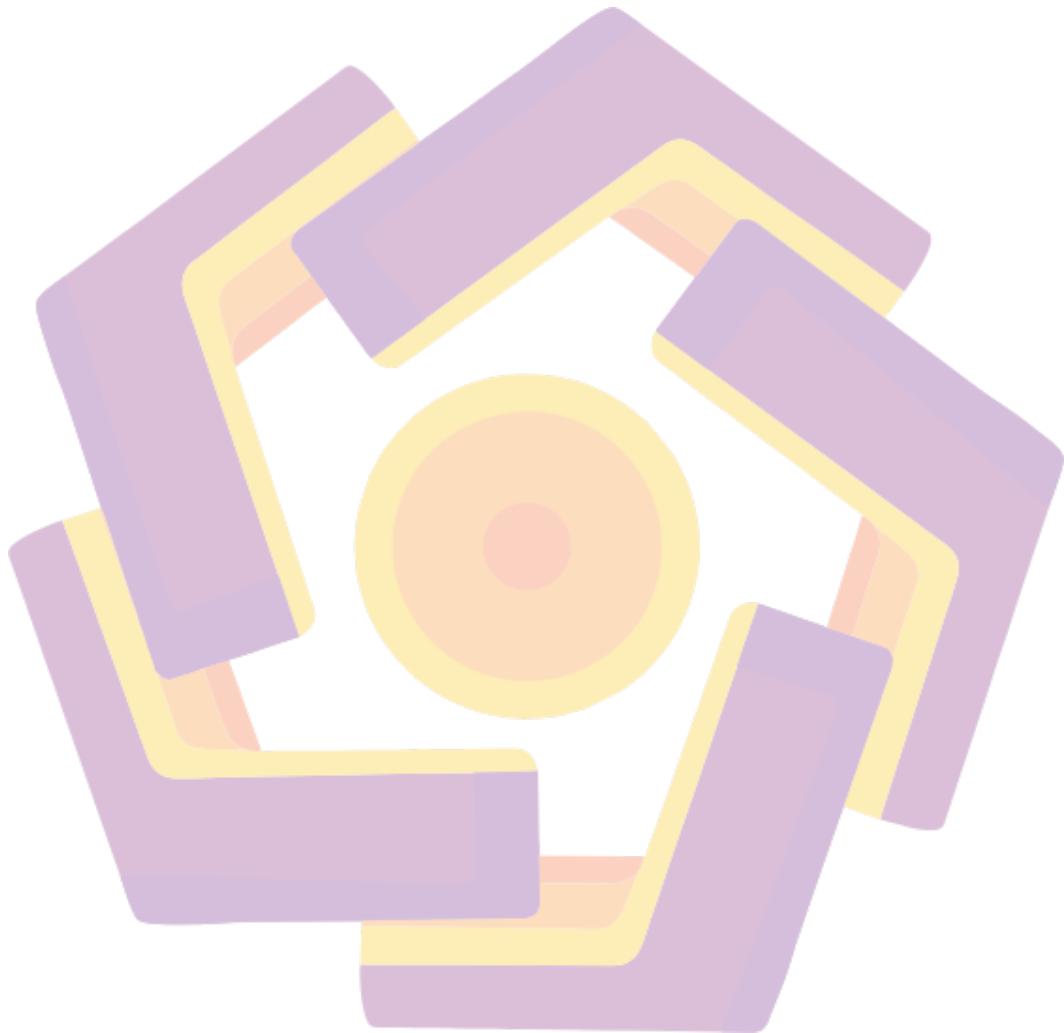
Penulis

## DAFTAR ISI

|                                     |       |
|-------------------------------------|-------|
| HALAMAN JUDUL                       | i     |
| HALAMAN PERSETUJUAN                 | ii    |
| HALAMAN PENGESAHAN                  | iii   |
| HALAMAN PERNYATAAN KEASLIAN SKRIPSI | iv    |
| HALAMAN PERSEMBAHAN                 | v     |
| KATA PENGANTAR                      | vi    |
| DAFTAR ISI                          | vii   |
| DAFTAR TABEL                        | xi    |
| DAFTAR GAMBAR                       | xii   |
| DAFTAR LAMPIRAN                     | xiii  |
| DAFTAR LAMBANG DAN SINGKATAN        | xiv   |
| DAFTAR ISTILAH                      | xvi   |
| INTISARI                            | xviii |
| <i>ABSTRACT</i>                     | xix   |
| BAB I PENDAHULUAN                   | 1     |
| 1.1 Latar Belakang                  | 1     |
| 1.2 Rumusan Masalah                 | 1     |
| 1.3 Batasan Masalah                 | 1     |
| 1.4 Tujuan Penelitian               | 1     |
| 1.5 Manfaat Penelitian              | 2     |
| 1.6 Sistematika Penulisan           | 2     |
| BAB II TINJAUAN PUSTAKA             | 2     |

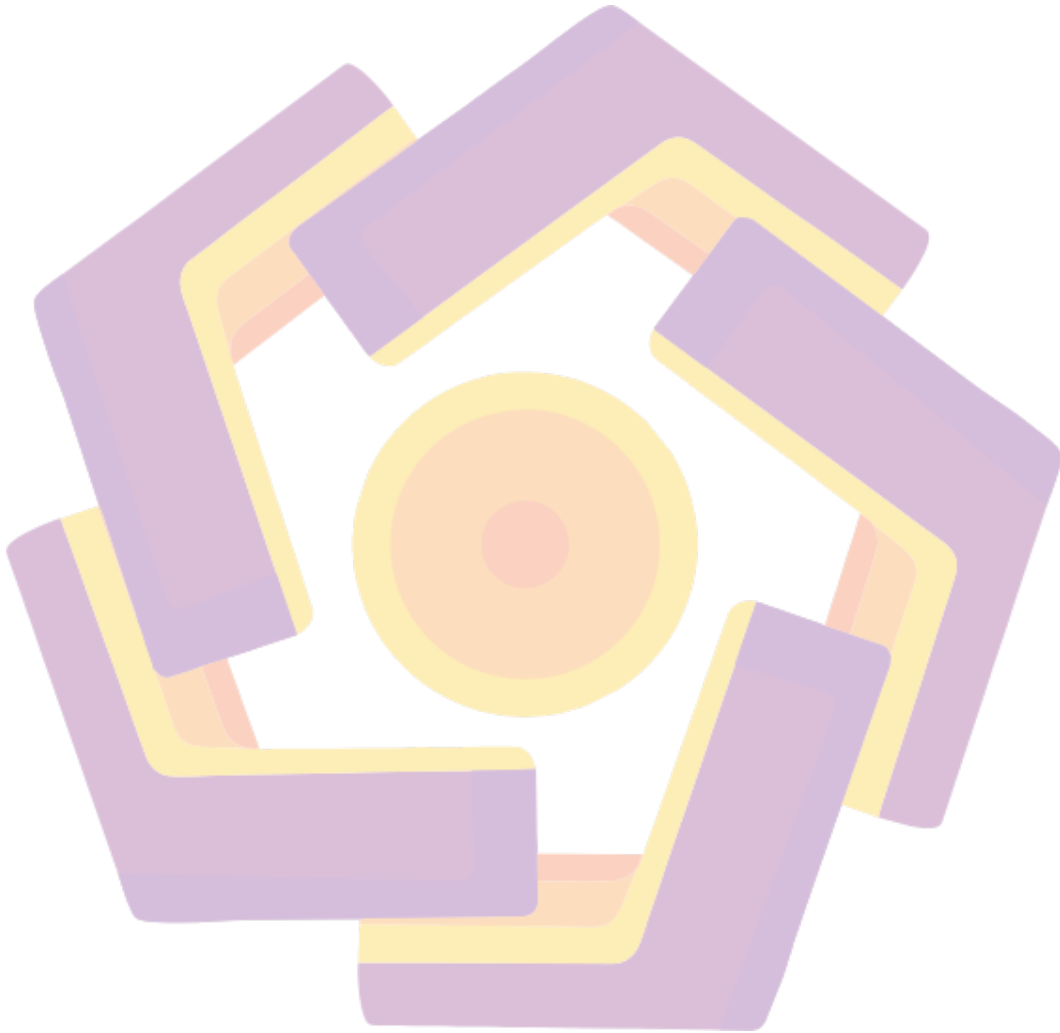
|                                  |    |
|----------------------------------|----|
| 2.1 Studi Literatur              | 2  |
| 2.2 Dasar Teori                  | 9  |
| 2.2.1 Penyakit Jantung           | 9  |
| 2.2.2 Machine Learning           | 9  |
| 2.2.3 Cleaning Data              | 9  |
| 2.2.4 Selection Feature          | 9  |
| 2.2.5 Standard Scaler            | 10 |
| 2.2.6 Model Random Forest        | 10 |
| 2.2.7 Model Logistic Regression  | 10 |
| 2.2.8 Model XGBoost              | 11 |
| 2.2.9 Hyper Parameter GridSearch | 11 |
| 2.2.10 Matix ROC-AUC             | 12 |
| <b>BAB III METODE PENELITIAN</b> | 13 |
| 3.1 Objek Penelitian             | 13 |
| 3.2 Alur Penelitian              | 14 |
| 3.2.1 Pengumpulan data           | 15 |
| 3.2.2 Cleaning data              | 15 |
| 3.2.3 Preprocessing              | 16 |
| 3.2.4 Splitting Data             | 16 |
| 3.2.5 Pemodelan                  | 16 |
| 3.2.6 Evaluasi Awal              | 16 |
| 3.2.7 Optimalisasi               | 17 |

|                                               |    |
|-----------------------------------------------|----|
| 3.2.8 Evaluasi Model                          | 17 |
| 3.3 Alat dan Bahan                            | 17 |
| BAB IV HASIL DAN PEMBAHASAN                   | 19 |
| 4.1 Hasil Pengolahan Data                     | 19 |
| 4.1.1 Cleaning Data                           | 19 |
| 4.1.2 Memeriksa Missing Value                 | 20 |
| 4.1.3 Distribusi Target                       | 21 |
| 4.1.4 Scoring Fitur                           | 22 |
| 4.1.5 Seleksi Fitur                           | 23 |
| 4.1.6 Standarisasi                            | 24 |
| 4.1.7 Splitting Data                          | 24 |
| 4.2 Pemodelan dan Optimalisasi hyperparameter | 25 |
| 4.2.1 Random Forest                           | 26 |
| 4.2.2 Logistic Regression                     | 27 |
| 4.2.3 XGBoost                                 | 30 |
| 4.3 Evaluasi Model                            | 32 |
| 4.3.1 Confusion Matrix                        | 33 |
| 4.3.2 ROC Curve                               | 35 |
| 4.4 Pembahasan                                | 35 |
| BAB V PENUTUP                                 | 37 |
| 5.1 Kesimpulan                                | 37 |
| 5.2 Saran                                     | 38 |



## DAFTAR TABEL

|                                                                                           |    |
|-------------------------------------------------------------------------------------------|----|
| Tabel 2. 1 Keaslian Penelitian .....                                                      | 5  |
| Tabel 3. 1 Keterangan Kolom Dataset .....                                                 | 15 |
| Tabel 4. 1 Perbandingan Performa Model Sebelum dan Setelah Hyperparameter<br>Tuning ..... | 32 |

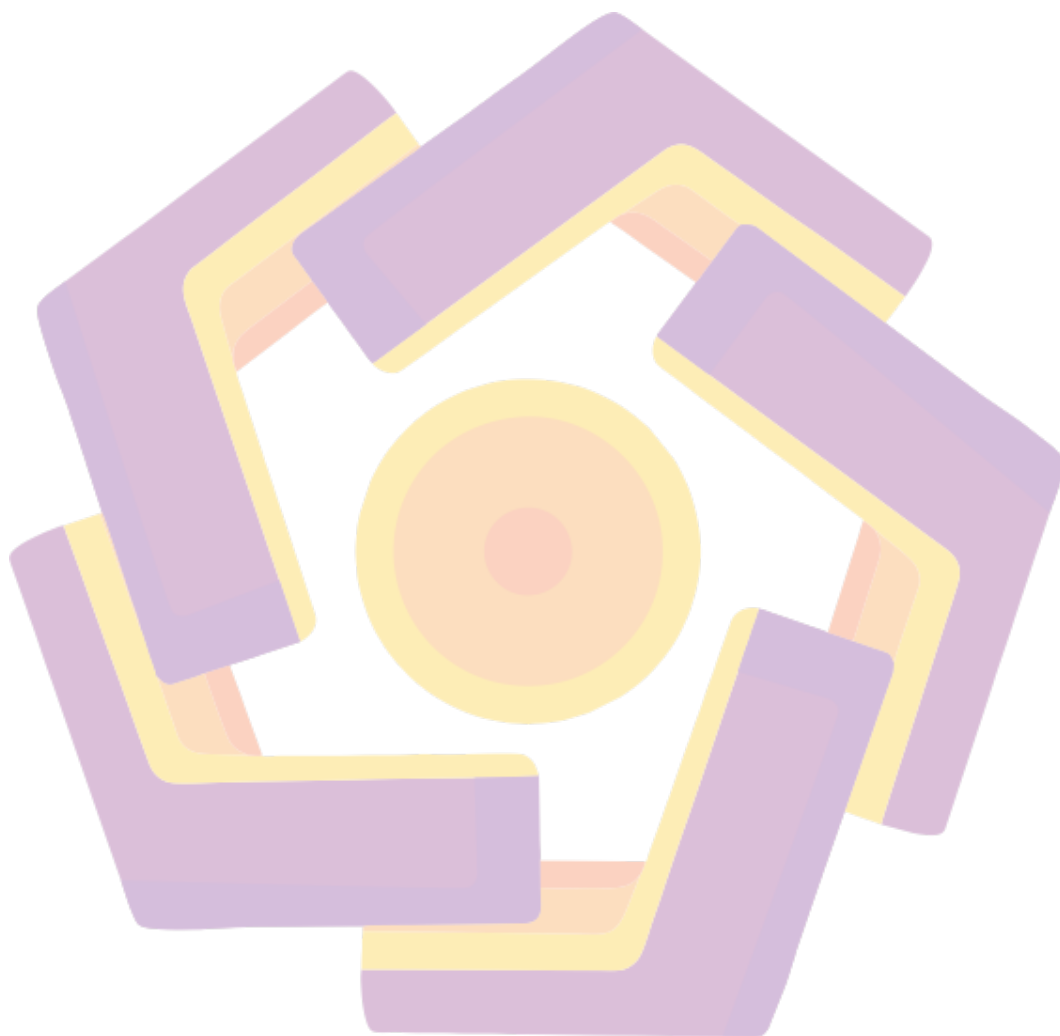


## DAFTAR GAMBAR

|                                                             |    |
|-------------------------------------------------------------|----|
| Gambar 4. 1 Dataset.....                                    | 19 |
| Gambar 4. 2 Cleaning Dataset .....                          | 20 |
| Gambar 4. 3 Missing Value .....                             | 20 |
| Gambar 4. 4 Distribusi Target .....                         | 21 |
| Gambar 4. 5 Fitur Score.....                                | 22 |
| Gambar 4. 6 Fitur Seleksi .....                             | 23 |
| Gambar 4. 7 Standarisasi .....                              | 24 |
| Gambar 4. 8 Splitting Data .....                            | 25 |
| Gambar 4. 9 Parameter random Forest .....                   | 26 |
| Gambar 4. 10 Optimalisasi RF dengan GridSearchCV .....      | 27 |
| Gambar 4. 11 Parameter Logistic Regression.....             | 28 |
| Gambar 4. 12 Optimalisasi LR dengan GridSearchCV .....      | 29 |
| Gambar 4. 13 Parameter XGBoost .....                        | 30 |
| Gambar 4. 14 Optimalisasi XGBoost dengan GridSearchCV ..... | 31 |
| Gambar 4. 15 Hasil confusion matrix .....                   | 34 |
| Gambar 4. 16 Hasil ROC Curve .....                          | 35 |

## DAFTAR LAMPIRAN

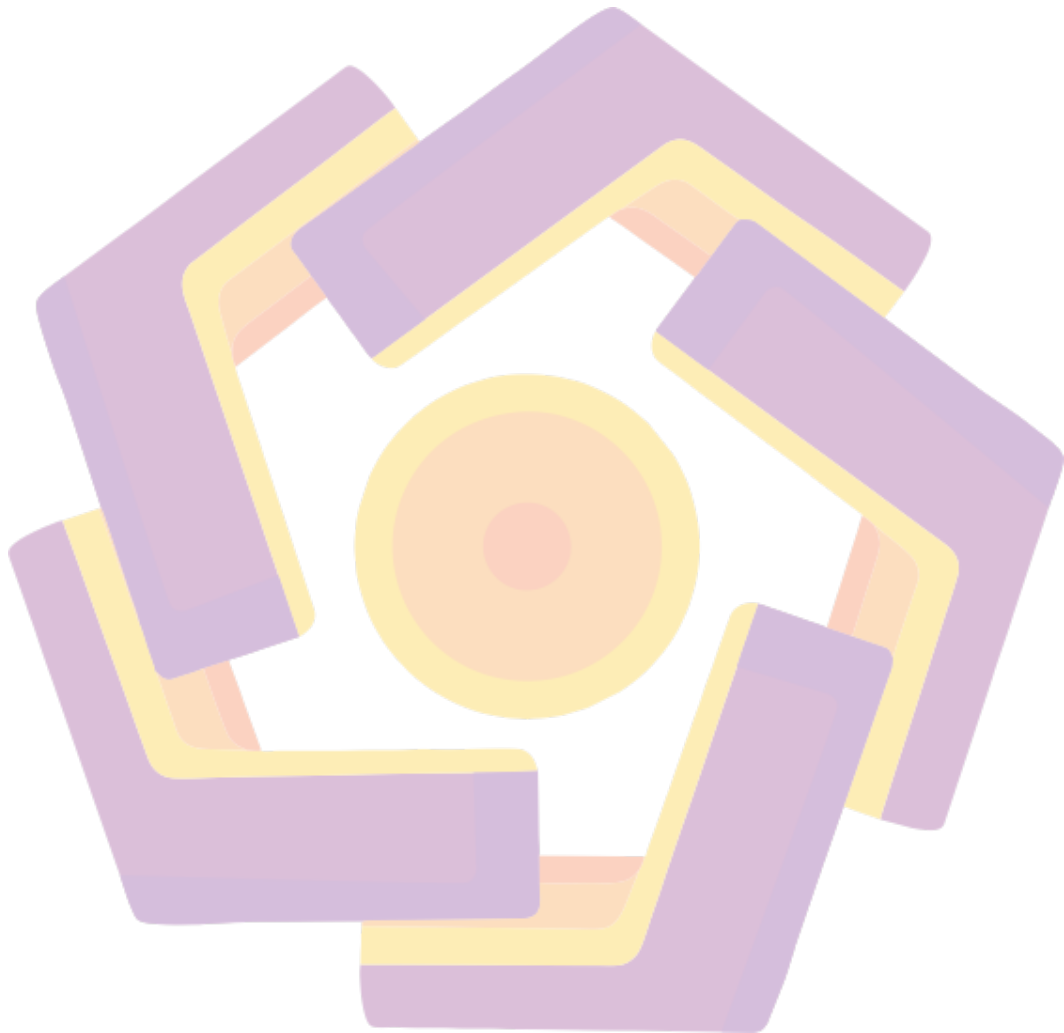
|                                     |    |
|-------------------------------------|----|
| Lampiran 1. Profil obyek Penelitian | 10 |
| Lampiran 2. Dokumentasi Penelitian  | 11 |



## DAFTAR LAMBANG DAN SINGKATAN

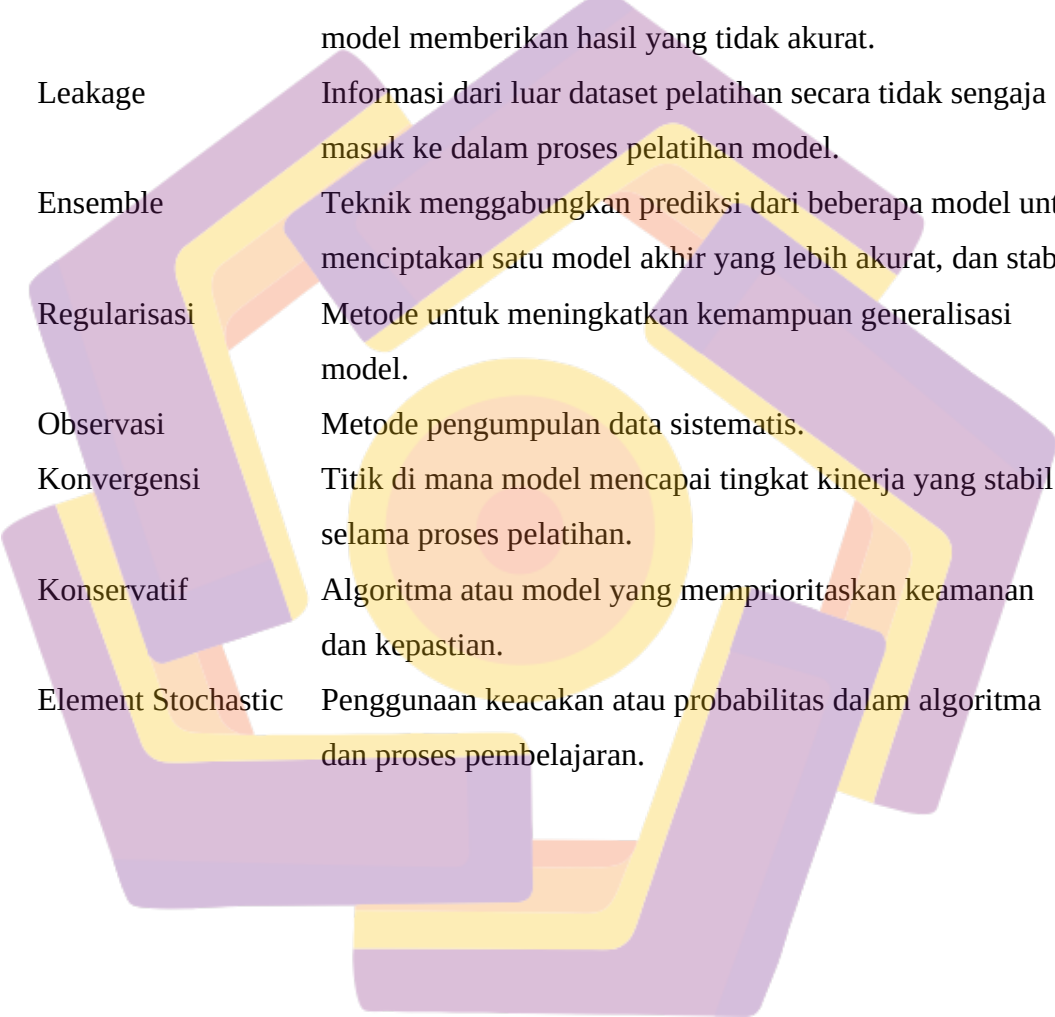
|                                    |                                                                   |
|------------------------------------|-------------------------------------------------------------------|
| LR                                 | Tahanan Listrik                                                   |
| K-NN                               | K-Nearest Neighbor                                                |
| SVM                                | Support Vector Machine                                            |
| DT                                 | Decision Tree                                                     |
| RF                                 | Random Forest                                                     |
| GS                                 | GridSearch                                                        |
| GA-FS                              | Genetic Algorithm Feature Selection                               |
| ROC                                | Receiver Operating Characteristic                                 |
| AUC                                | Area Under the Curve                                              |
| TP                                 | True Positif                                                      |
| TN                                 | True Negatif                                                      |
| FP                                 | False Positif                                                     |
| FN                                 | False Negative                                                    |
| $\mu$                              | Rata-rata (mean) dari nilai fitur pada data training              |
| $\chi$                             | Nilai asli dari fitur numerik sebelum distandarisasi              |
| $Z$                                | Nilai fitur setelah distandarisasi                                |
| $\sigma$                           | Standar deviasi dari nilai fitur pada data training               |
| $\pi(x)$                           | Probabilitas bahwa observasi termasuk dalam kelas 1               |
| exp                                | Fungsi eksponensial                                               |
| $\beta_0$                          | Koefisien intersep                                                |
| $\beta_1, \beta_2, \dots, \beta_P$ | Koefisien regresi (bobot) untuk masing-masing variabel independen |
| $P$                                | Jumlah variabel independen (fitur) yang digunakan dalam model     |
| $L$                                | Nilai total fungsi loss                                           |
| $y_i$                              | Nilai target sebenarnya untuk sampel ke-i                         |
| $p_i$                              | Probabilitas prediksi model untuk sampel ke-i termasuk kelas 1    |
| $n$                                | Jumlah total sampel dalam dataset training                        |
| $\hat{y}_i$                        | Nilai prediksi akhir untuk sampel ke-i                            |
| $f_k(X_i)$                         | Fungsi ke-k yang memberikan kontribusi skor pada sampel $x_i$     |

|          |                                                                                  |
|----------|----------------------------------------------------------------------------------|
| $x_i$    | Vektor fitur (input) untuk sampel ke-i                                           |
| $k$      | Jumlah total pohon keputusan yang dibangun secara sekuensial dalam model XGBoost |
| $\Sigma$ | Simbol penjumlahan                                                               |



## DAFTAR ISTILAH

|                |                                                                                                                                                                     |
|----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Hyperparameter | Variabel konfigurasi eksternal yang diatur secara manual sebelum melatih model <i>machine learning</i> untuk mengontrol bagaimana model belajar                     |
| Tuning         | Proses menyesuaikan atau memodifikasi suatu sistem atau perangkat agar bekerja lebih optimal, efisien, atau sesuai dengan preferensi tertentu                       |
| GridSearchCV   | Metode penyetelan hyperparameter                                                                                                                                    |
| Performa       | Kinerja Tingkat keberhasilan atau hasil yang dicapai                                                                                                                |
| Teoritis       | Sesuatu yang berkaitan dengan teori                                                                                                                                 |
| Mengkaji       | Mempelajari, menelaah, memeriksa, atau menyelidiki                                                                                                                  |
| Ensemble       | Menggabungkan beberapa model individual (base learners) untuk membuat satu prediksi akhir yang lebih akurat dan stabil daripada model tunggal                       |
| Intervensi     | Tindakan campur tangan atau keterlibatan dalam suatu situasi, konflik, atau proses yang sedang berlangsung, untuk memperbaiki keadaan, atau mencegah dampak buruk.  |
| Diagnosis      | Proses ilmiah untuk mengidentifikasi jenis penyakit atau kondisi kesehatan yang dialami seseorang berdasarkan gejala.                                               |
| Aktual         | Merujuk pada sesuatu yang benar-benar ada, terjadi, atau sedang hangat dibicarakan karena masih baru.                                                               |
| Proporsi       | Prinsip seni dan desain yang merujuk pada perbandingan ukuran, kuantitas, atau derajat yang seimbang antara satu bagian dengan bagian lain atau dengan keseluruhan. |
| Pradigma       | Kerangka berpikir, model, atau pola pandang mendasar yang digunakan individu maupun kelompok untuk memahami realitas atau fenomena tertentu                         |
| Subset         | Kumpulan elemen yang seluruh anggotanya merupakan bagian dari himpunan lain yang lebih besar.                                                                       |



|                    |                                                                                                                     |
|--------------------|---------------------------------------------------------------------------------------------------------------------|
| Integritas         | Data dan model tetap akurat, konsisten, dan tidak berubah.                                                          |
| Bias               | Perbedaan antara rata rata hasil prediksi dari model ML dengan data nilai yang sebenarnya.                          |
| Interpretasi       | Pemberian tafsiran, kesan, pandangan teoretis, atau makna mendalam terhadap suatu objek.                            |
| Distorsi           | Penyimpangan atau manipulasi data yang menyebabkan model memberikan hasil yang tidak akurat.                        |
| Leakage            | Informasi dari luar dataset pelatihan secara tidak sengaja masuk ke dalam proses pelatihan model.                   |
| Ensemble           | Teknik menggabungkan prediksi dari beberapa model untuk menciptakan satu model akhir yang lebih akurat, dan stabil. |
| Regularisasi       | Metode untuk meningkatkan kemampuan generalisasi model.                                                             |
| Observasi          | Metode pengumpulan data sistematis.                                                                                 |
| Konvergensi        | Titik di mana model mencapai tingkat kinerja yang stabil selama proses pelatihan.                                   |
| Konservatif        | Algoritma atau model yang memprioritaskan keamanan dan kepastian.                                                   |
| Element Stochastic | Penggunaan keacakan atau probabilitas dalam algoritma dan proses pembelajaran.                                      |

## INTISARI

Penyakit jantung merupakan penyebab utama kematian global, dengan proyeksi WHO menunjukkan sekitar 20,5 juta kematian pada 2025 dan akan meningkat hingga 35,6 juta pada 2050 akibat faktor risiko seperti hipertensi, pola makan buruk, dan kurang olahraga. Di Indonesia, prevalensi penyakit jantung terus meningkat, dengan data BPJS Kesehatan mencatat beban biaya katastropik tertinggi mencapai puluhan triliun rupiah per tahun, serta tren naik dari 0,5% (2013) menjadi 1,5% (2018) berdasarkan Riskesdas. Tantangan diagnosis dini dan analisis data medis besar mendorong pemanfaatan machine learning untuk prediksi akurat.

Penelitian ini membandingkan performa tiga algoritma machine learning Logistic Regression, Random Forest, dan XGBoost dalam klasifikasi penyakit jantung menggunakan dataset publik dari Kaggle dengan 1048 baris. Proses meliputi cleaning data, seleksi fitur dengan SelectKBest (menghapus kolesterol karena skor rendah), standarisasi, splitting data (80:20 dengan stratify), pemodelan, serta optimalisasi hyperparameter menggunakan GridSearchCV.

Hasil menunjukkan bahwa Random Forest dan XGBoost mencapai nilai ROC-AUC tertinggi 0.90 setelah tuning, dengan peningkatan signifikan dibandingkan tanpa tuning, sementara Logistic Regression berada di 0.84. Fitur paling berpengaruh adalah ST slope, chest pain type, dan exercise angina. Random Forest direkomendasikan sebagai model terbaik karena keseimbangan akurasi, stabilitas, dan kemampuan diskriminasi. Penelitian ini memberikan kontribusi untuk deteksi dini penyakit jantung berbasis data publik, dengan saran pengembangan menggunakan dataset lokal Indonesia dan integrasi ke sistem kesehatan.

**Kata kunci:** GridSearchCV, Machine Learning, Random Forest, Logistic Regression, XGBoost.

## **ABSTRACT**

*Heart disease is the leading cause of global death, with WHO projections indicating approximately 20.5 million deaths in 2025 and rising to 35.6 million by 2050 due to risk factors such as hypertension, poor diet, and lack of exercise. In Indonesia, the prevalence of heart disease continues to rise, with BPJS health data recording the highest catastrophic cost burden reaching tens of trillions of rupiah per year, and an upward trend from 0.5% (2013) to 1.5% (2018) based on RISKESDAS. The challenges of early diagnosis and big medical data analysis encourage the use of machine learning for accurate predictions.*

*This study compares the performance of three machine learning algorithms Logistic Regression, Random Forest, and XGBoost in classifying heart disease using a public dataset from Kaggle with 1048 rows. The process includes data cleaning, feature selection with SelectKBest (removing cholesterol due to its low score), standardization, data splitting (80:20 with stratify), modeling, and hyperparameter optimization using GridSearchCV.*

*The results show that Random Forest and XGBoost achieved the highest ROC-AUC value of 0.90 after tuning, with a significant improvement compared to without tuning, while Logistic Regression was at 0.84. The most influential features were ST slope, chest pain type, and exercise angina. Random Forest was recommended as the best model due to its balance of accuracy, stability, and discrimination ability. This study contributes to early detection of heart disease based on public data, with suggestions for development using local Indonesian datasets and integration into the health system.*

**Keyword:** GridSearchCV, Machine Learning, Random Forest, Logistic Regression, XGBoost.