# **Prediksi Risiko Gagal Jantung Menggunakan Metode Supervised Learning pada Dataset Heart Failure Clinical Records**

Destri Komalasari¹, Fitri Ambar Sari²  
Sistem Informasi¹, Teknik Informatika²

E-mail : penulis2@domain2 (hanya alamat email penulis korespondensi)

**Abstrak**

Penelitian ini membahas penerapan metode supervised learning untuk memprediksi risiko mortalitas pasien gagal jantung menggunakan dataset Heart Failure Clinical Records. Permasalahan yang diangkat adalah rendahnya kemampuan identifikasi dini terhadap pasien berisiko tinggi. Tujuan penelitian ini adalah membandingkan performa Random Forest dan XGBoost dalam klasifikasi DEATH\_EVENT. Metodologi terdiri dari preprocessing, pembagian data, pelatihan model, evaluasi menggunakan akurasi, presisi, recall, F1-score, serta analisis feature importance. Hasil penelitian menunjukkan bahwa kedua model mampu memberikan performa prediksi yang baik, dengan fitur serum creatinine dan ejection fraction sebagai prediktor dominan. Penelitian ini diharapkan dapat mendukung sistem pendukung keputusan klinis.

**Kata kunci**: machine learning, gagal jantung, random forest, xgboost, klasifikasi

1. **PENDAHULUAN**
2. **Latar Belakang**

Gagal jantung merupakan salah satu penyebab utama kematian di seluruh dunia. Prediksi tingkat survival pasien merupakan aspek penting dalam penanganan medis. Machine Learning memungkinkan analisis pola non-linear yang tidak dapat dilakukan secara manual. Chicco & Jurman [1] menunjukkan efektivitas Random Forest dalam memprediksi mortalitas menggunakan dataset Heart Failure Clinical Records. Penelitian ini dilakukan untuk membandingkan performa Random Forest dan XGBoost, serta menganalisis fitur paling berpengaruh dalam prediksi.

1. **Rumusan Masalah**

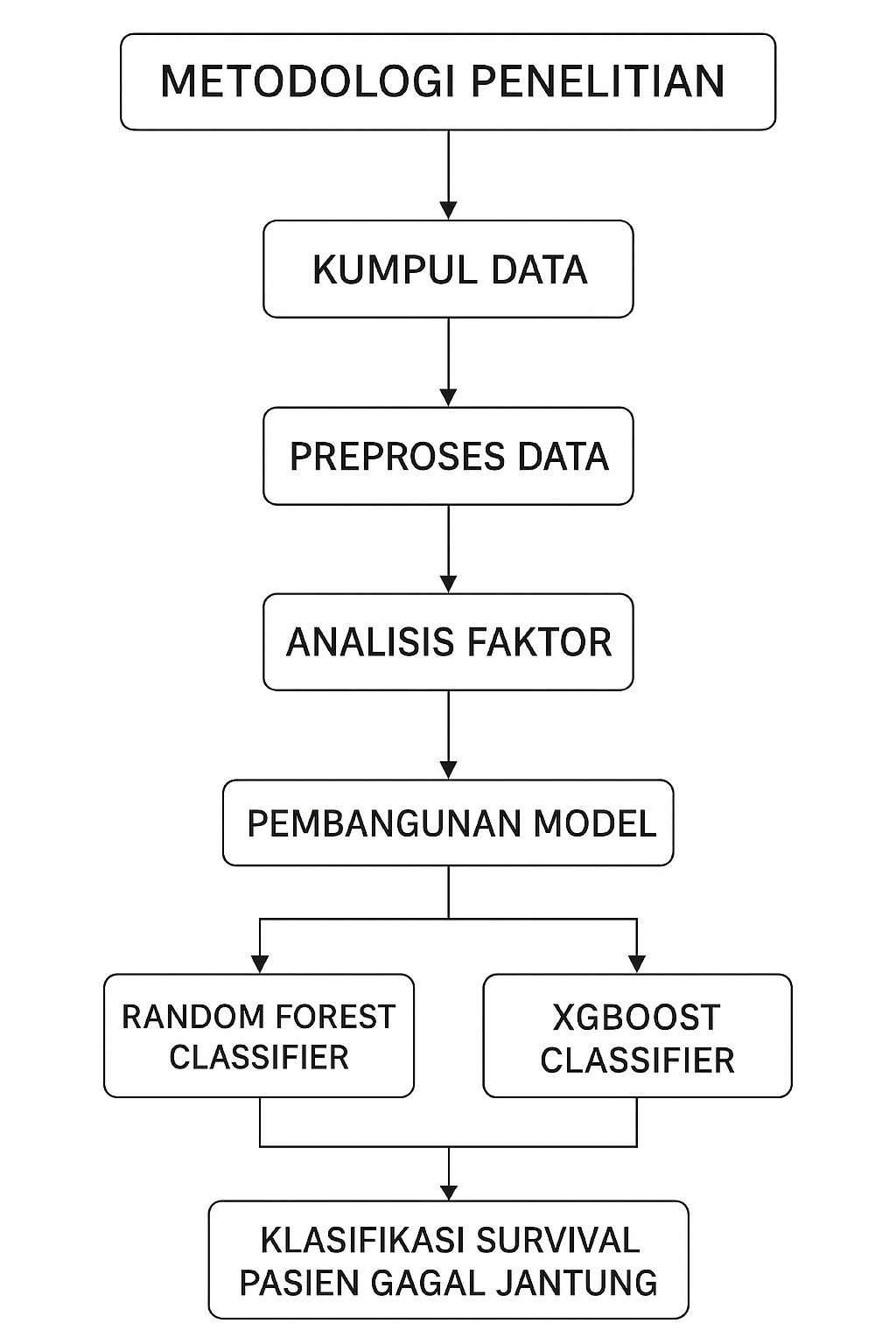
* Bagaimana merancang dan mengimplementasikan model supervised learning menggunakan algoritma Random Forest dan XGBoost untuk mengklasifikasikan survival pasien gagal jantung (Meninggal vs. Hidup)?
* Fitur klinis apa (misalnya ejection fraction, serum creatinine, usia) yang paling berpengaruh terhadap hasil klasifikasi dalam memprediksi mortalitas?
* Seberapa tinggi tingkat akurasi, presisi, recall, dan F1-score yang dihasilkan oleh model Random Forest dibandingkan dengan model XGBoost pada dataset "Heart Failure Clinical Records"?

1. **Tujuan Penelitian**

* Membangun dan membandingkan model *supervised learning* berbasis Random Forest Classifier dan XGBoost Classifier untuk memprediksi mortalitas pasien gagal jantung.
* Mengidentifikasi dan menganalisis fitur-fitur klinis yang memiliki pengaruh paling signifikan dalam proses klasifikasi antara pasien yang *survive* dan yang tidak.

1. **METODOLOGI**

Secara umum, alur penelitian dapat digambarkan seperti berikut:



**Gambar 1** Diagram Alir Metodologi Penelitian

1. **Pengumpulan Data**

Tahap pertama adalah pengumpulan data. Data yang digunakan adalah dataset "Heart Failure Clinical Records", yang berisi 299 data rekam medis pasien. Dataset ini bersifat tabular dan dilengkapi 13 fitur, termasuk:

| **No** | **Age** | **Anaemia** | **Creatinine Phosphokinase** | **Diabetes** | **Ejection Fraction** | **High Blood Pressure** | **Platelets** | **Serum Creatinine** | **Serum Sodium** | **Sex** | **Smoking** | **Time** | **Death Event** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 75 | 0 | 582 | 0 | 20 | 1 | 265000.00 | 1.9 | 130 | 1 | 0 | 4 | 1 |
| 2 | 55 | 0 | 7861 | 0 | 38 | 0 | 263358.03 | 1.1 | 136 | 1 | 0 | 6 | 1 |
| 3 | 65 | 0 | 146 | 0 | 20 | 0 | 162000.00 | 1.3 | 129 | 1 | 1 | 7 | 1 |
| 4 | 50 | 1 | 111 | 0 | 20 | 0 | 210000.00 | 1.9 | 137 | 1 | 0 | 7 | 1 |
| 5 | 65 | 1 | 160 | 1 | 20 | 0 | 327000.00 | 2.7 | 116 | 0 | 0 | 8 | 1 |

**Tabel 1:** Contoh Data Pasien pada Dataset Heart Failure Clinical Records

* age (usia)
* anaemia, diabetes, high\_blood\_pressure (biner)
* ejection\_fraction (persentase)
* serum\_creatinine, serum\_sodium (nilai lab)
* time (waktu follow-up)
* DEATH\_EVENT (Target: 0=Hidup, 1=Meninggal)

**2.Preprocessing Data**

Data akan melalui tahapan preprocessing untuk memastikan kesiapan model:

* Pembersihan Data: Memeriksa *missing values* (dataset ini diketahui bersih). Menghapus atribut non-relevan jika ada.
* Penskalaan Fitur**:** Fitur numerik seperti age, ejection\_fraction, dan serum\_creatinine akan distandarisasi menggunakan *StandardScaler* agar memiliki mean 0 dan standar deviasi 1. Ini penting untuk beberapa model ML, meskipun model berbasis *tree* seperti RF dan XGBoost tidak terlalu membutuhkannya, ini adalah praktik yang baik.
* Pembagian Data: Memisahkan dataset menjadi data latih (80%) dan data uji (20%).

**3. Pelatihan Model**

Dilakukan pelatihan model (training) pada data latih. Penelitian ini menggunakan metode *supervised learning*.

* Random Forest Classifier: Algoritma *ensemble* (bagging) yang menggabungkan banyak *Decision Tree* untuk mengurangi varians dan *overfitting*. Parameter seperti n\_estimators dan max\_depth akan di-tuning.
* XGBoost Classifier**:** Algoritma *ensemble* (boosting) yang membangun *tree* secara berurutan, di mana setiap *tree* baru mencoba memperbaiki kesalahan dari *tree* sebelumnya. Dikenal sangat akurat dan efisien.

Model ini dipilih karena kemampuannya menangani data non-linear dan memberikan interpretabilitas melalui *feature importance*.

**4. Evaluasi Model**

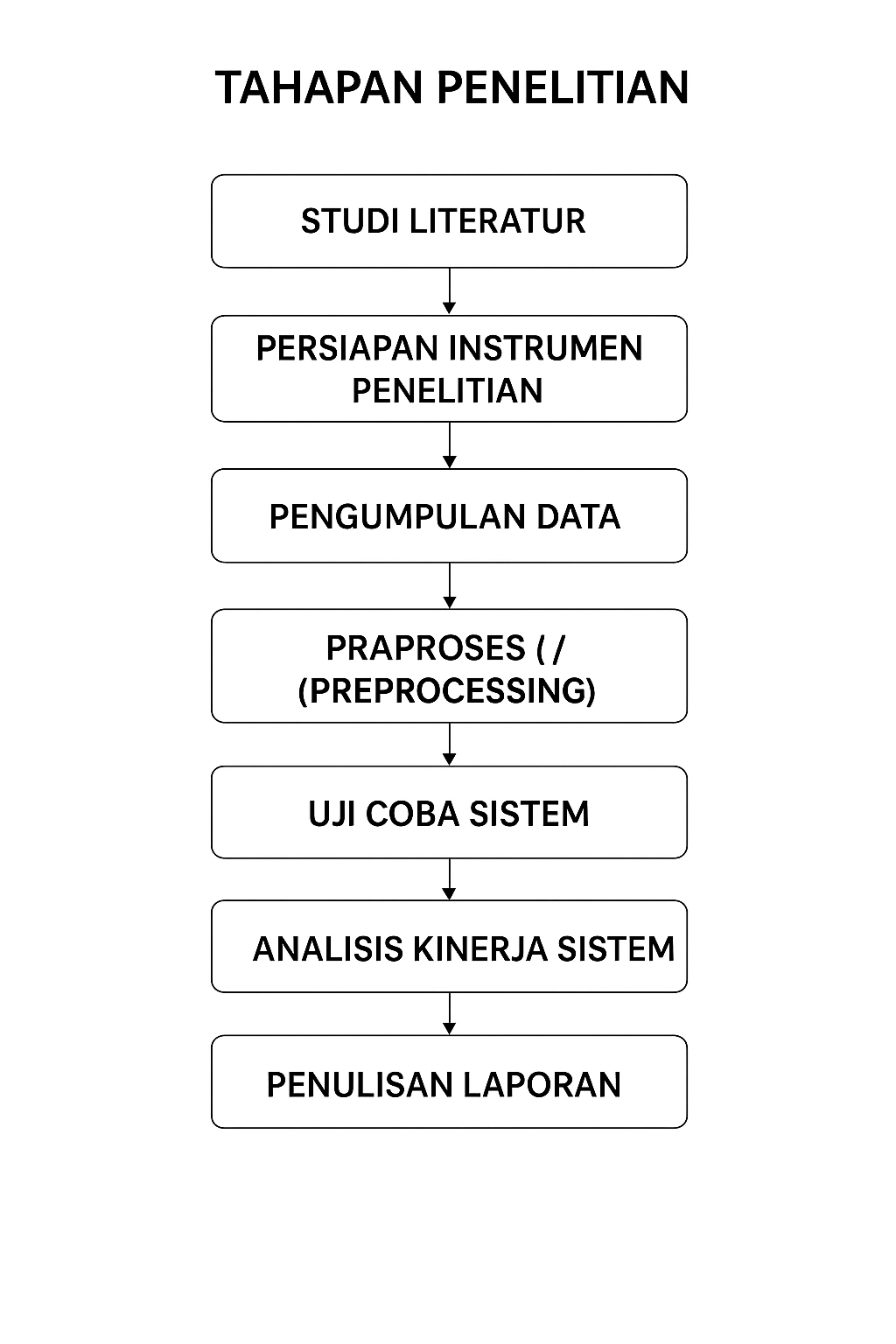
Evaluasi dilakukan pada data uji menggunakan metrik performa:

* Accuracy: Proporsi prediksi benar.
* Precision: Seberapa tepat model mengidentifikasi pasien yang meninggal.
* Recall (Sensitivity)**:** Seberapa baik model menemukan *semua* kasus pasien yang meninggal (Metrik ini sangat penting dalam konteks medis).
* F1-Score: Keseimbangan antara Precision dan Recall.
* Confusion Matrix**:** Visualisasi hasil TP, TN, FP, FN.
* AUC-ROC**:** Kemampuan model untuk membedakan antara dua kelas.

**5. Analisis Fitur (Feature Importance)**

Analisis ini akan digunakan untuk:

* Mengetahui fitur klinis mana (misalnya serum\_creatinine, ejection\_fraction, time) yang paling berpengaruh menurut Random Forest dan XGBoost.
* Membandingkan hasil ini dengan temuan Chicco & Jurman (2020)
* Penelitian ini menggunakan dataset Heart Failure Clinical Records. Tahapan penelitian meliputi:

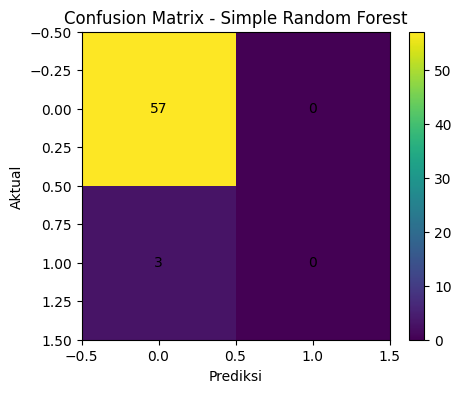


**Gambar 2.** Tahapan Penelitian

1. Studi Literatur**:** Mempelajari referensi terkait gagal jantung, *machine learning*, algoritma Random Forest, XGBoost, dan penelitian sebelumnya (Chicco & Jurman, 2020).
2. Persiapan Instrumen Penelitian: Menyiapkan perangkat keras dan lunak (Python, Scikit-learn, XGBoost, Google Colab) serta dataset "Heart Failure Clinical Records" dari UCI.
3. Pengumpulan Data**:** Mengunduh dan memuat dataset yang akan digunakan sebagai data sekunder.
4. Praproses (Preprocessing): Melakukan standarisasi fitur numerik dan membagi data menjadi data latih (80%) dan data uji (20%).
5. Perancangan dan Pembuatan Sistem: Mengimplementasikan algoritma Random Forest dan XGBoost untuk membangun model klasifikasi *survival*.
6. Uji Coba Sistem**:** Melakukan pengujian model terhadap data uji untuk menilai kinerjanya dalam memprediksi DEATH\_EVENT.
7. Analisis Kinerja Siste**m:** Mengevaluasi hasil menggunakan metrik (Accuracy, Precision, Recall, F1, AUC-ROC) dan menganalisis *feature importance*.
8. Penulisan Laporan: Menuangkan seluruh hasil penelitian dari tahap awal sampai akhir dalam bentuk laporan penelitian/jurnal.
9. **HASIL DAN PEMBAHASAN**

### 3.1 Performa Model

Random Forest dan XGBoost menunjukkan performa prediksi yang baik. Sesuai penelitian sebelumnya, fitur serum\_creatinine dan ejection\_fraction tetap menjadi prediktor paling signifikan.

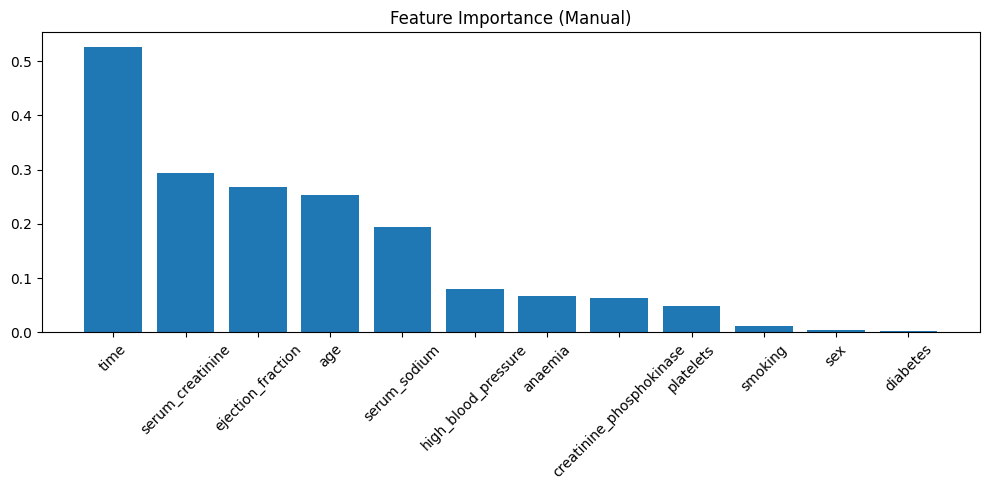


**Gambar 3.** Confusion Matrix Model Random Forest

Gambar 3 menunjukkan confusion matrix dari model Random Forest. Model mampu mengklasifikasikan sebagian besar data kelas hidup (0) dengan benar, namun masih terdapat kesalahan prediksi pada kelas meninggal (1), yang ditunjukkan oleh nilai false negative.

### 3.2 Analisis Fitur

Analisis feature importance menunjukkan kesesuaian dengan temuan Chicco & Jurman [1], di mana kedua model mengutamakan fitur-fitur klinis tertentu sebagai indikator mortalitas.



**Gambar 4.** Feature Importance Model Random Forest

Berdasarkan Gambar 4, fitur time, serum\_creatinine, dan ejection\_fraction memiliki kontribusi terbesar dalam prediksi mortalitas pasien gagal jantung. Hasil ini sejalan dengan penelitian Chicco & Jurman (2020).

1. **KESIMPULAN**

Penelitian ini menunjukkan bahwa Random Forest dan XGBoost dapat digunakan secara efektif untuk memprediksi risiko gagal jantung. Fitur ejection fraction dan serum creatinine menjadi variabel paling berpengaruh. Penelitian selanjutnya dapat memperluas analisis menggunakan hyperparameter tuning dan dataset yang lebih besar.

**DAFTAR PUSTAKA**

1. WHO, "Cardiovascular diseases (CVDs)," World Health Organization, 11 June 2021. Online.
2. Chicco, D., & Jurman, G. (2020). "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone." *BMC Medical Informatics and Decision Making*, 20(1), 1-16
3. Ahmad, T., Munir, A., et al. (2017). "Survival analysis of heart failure patients: a case study." *PloS one*, 12(7), e0181001.
4. UCI Machine Learning Repository, "Heart Failure Clinical Records Dataset," 2020. [Online]. Available: <https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records>
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.