

**HYBRID PREDICTION MODEL PENYAKIT JANTUNG
MENGUNAKAN ALGORITMA RANDOM FOREST, XGBOOST
DAN LINEAR AGRESSION**

LAPORAN RISET INFORMATIKA



Oleh :

DAFFA TUNGGGA WISESA

NPM 21081010243

**PROGRAM STUDI INFORMATIKA FAKULTAS ILMU
KOMPUTER UNIVERSITAS PEMBANGUNAN
NASIONAL "VETERAN" JAWA TIMUR**

2024

BAB 1: PENDAHULUAN

1.1 Latar Belakang Penyakit jantung merupakan salah satu penyebab utama kematian di seluruh dunia. Deteksi dini terhadap penyakit ini sangat penting untuk mengurangi risiko fatal. Dalam bidang ilmu data dan kecerdasan buatan, berbagai pendekatan berbasis machine learning telah digunakan untuk memprediksi kemungkinan seseorang menderita penyakit jantung. Algoritma seperti Random Forest, XGBoost, dan Logistic Regression merupakan model yang sering digunakan karena efektivitasnya dalam analisis data medis.

Penelitian ini bertujuan untuk mengevaluasi dan meningkatkan performa prediksi penyakit jantung menggunakan metode ensemble learning, khususnya Stacking Classifier, yang menggabungkan beberapa model dasar untuk meningkatkan akurasi dan keandalan prediksi.

1.2 Rumusan Masalah

1. Bagaimana performa model Stacking Classifier dalam memprediksi penyakit jantung dibandingkan dengan pendekatan model individu?
2. Seberapa baik sensitivitas, spesifisitas, dan akurasi model dalam mendeteksi penyakit jantung?
3. Bagaimana visualisasi metrik evaluasi, seperti Confusion Matrix dan ROC Curve, dapat memberikan wawasan tentang performa model?

1.3 Tujuan Penelitian

1. Membangun model prediksi penyakit jantung menggunakan Stacking Classifier dengan Random Forest, XGBoost, dan Logistic Regression sebagai komponen modelnya.
2. Mengevaluasi performa model menggunakan metrik akurasi, sensitivitas, spesifisitas, dan ROC AUC.
3. Menyediakan visualisasi metrik evaluasi untuk interpretasi yang lebih baik.

1.4 Manfaat Penelitian Penelitian ini diharapkan dapat memberikan kontribusi pada pengembangan sistem pendukung keputusan medis yang lebih akurat untuk membantu dokter dan praktisi medis dalam diagnosis penyakit jantung.

BAB 2: TINJAUAN PUSTAKA

2.1 Penyakit Jantung dan Deteksi Dini Penyakit jantung adalah gangguan pada jantung yang memengaruhi fungsinya. Deteksi dini terhadap penyakit ini memainkan peran penting dalam meningkatkan harapan hidup pasien. Dalam beberapa dekade terakhir, machine learning telah menjadi salah satu pendekatan utama dalam pengembangan alat bantu diagnostik untuk penyakit jantung.

2.2 Machine Learning dalam Prediksi Penyakit Jantung Beberapa algoritma machine learning yang umum digunakan dalam prediksi penyakit jantung adalah:

- **Random Forest:** Model berbasis ensemble yang menggunakan banyak pohon keputusan untuk meningkatkan akurasi.
- **XGBoost:** Algoritma boosting yang terkenal dengan performanya yang tinggi dalam berbagai kompetisi machine learning.
- **Logistic Regression:** Model yang sederhana namun efektif untuk tugas klasifikasi biner.

2.3 Ensemble Learning dan Stacking Classifier Ensemble learning adalah pendekatan yang menggabungkan beberapa model untuk meningkatkan akurasi prediksi. Salah satu metode ensemble adalah Stacking Classifier, di mana model dasar (base models) menghasilkan prediksi yang kemudian digunakan oleh model meta (meta-model) untuk membuat keputusan akhir. Kombinasi model ini dapat mengatasi kelemahan dari masing-masing model individu.

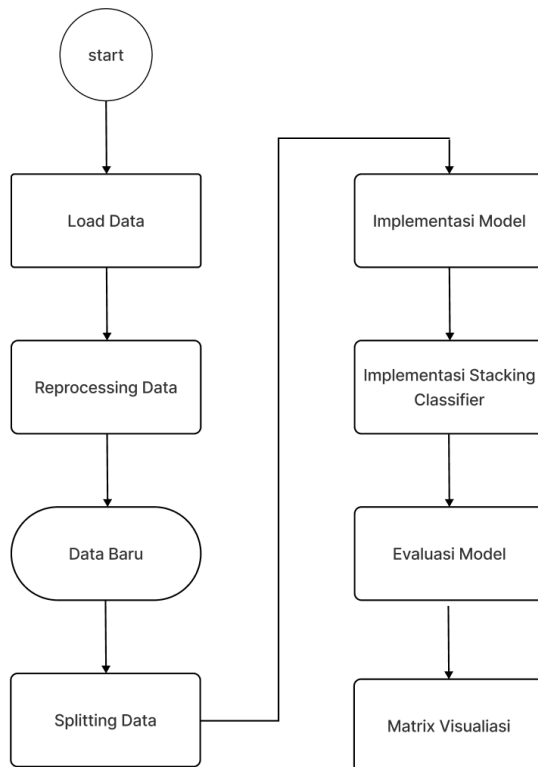
2.4 Evaluasi Model Machine Learning Metrik yang umum digunakan dalam evaluasi model prediksi adalah:

- **Akurasi:** Proporsi prediksi yang benar terhadap total prediksi.
- **Sensitivitas (Recall):** Kemampuan model untuk mendeteksi positif dengan benar.
- **Spesifisitas:** Kemampuan model untuk mendeteksi negatif dengan benar.

- **ROC AUC:** Ukuran yang menggabungkan sensitivitas dan spesifisitas pada berbagai ambang batas.

BAB 3: METODOLOGI

Bab ini menjelaskan secara rinci langkah-langkah yang diambil dalam penelitian ini untuk membangun model prediksi penyakit jantung. Dimulai dengan deskripsi dataset yang digunakan, proses preprocessing data, hingga pembagian data untuk pelatihan dan pengujian. Selanjutnya, dijelaskan model yang diterapkan, termasuk kombinasi beberapa algoritma dalam Stacking Classifier untuk meningkatkan akurasi prediksi. Proses implementasi model dijabarkan secara sistematis, mulai dari pemilihan model dasar hingga pelatihan meta-model. Terakhir, evaluasi model dilakukan menggunakan berbagai metrik untuk mengukur performa, serta visualisasi hasil yang mendukung interpretasi data dan prediksi. Dengan struktur yang terorganisir, bab ini bertujuan memberikan gambaran lengkap mengenai pendekatan yang digunakan dalam penelitian ini.




3.1 Dataset

Dataset yang digunakan adalah "heartdesease.data" dari UCI Healt Machine Learning yang berisi informasi pasien dengan fitur-fitur seperti usia, jenis kelamin, tekanan darah, kolesterol, dan lainnya. Dataset ini terdiri dari 14 kolom, di mana kolom terakhir adalah target variabel yang menunjukkan keberadaan penyakit jantung.

3.2 Preprocessing Data

proses preprocessing data dilakukan dengan membaca dataset menggunakan pustaka Pandas dan memberikan header pada kolom-kolomnya agar lebih mudah dikenali. Setelah itu, data disimpan dalam format CSV untuk mempermudah pengelolaan di tahap-tahap selanjutnya.

1 to 10 of 303 entries 

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	1
67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0
56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0
62.0	0.0	4.0	140.0	268.0	0.0	2.0	160.0	0.0	3.6	3.0	2.0	3.0	1
57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0	0.0	3.0	0
63.0	1.0	4.0	130.0	254.0	0.0	2.0	147.0	0.0	1.4	2.0	1.0	7.0	1
53.0	1.0	4.0	140.0	203.0	1.0	2.0	155.0	1.0	3.1	3.0	0.0	7.0	1

3.3 Pembagian Data

dataset dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih, yang merupakan 80% dari total dataset, digunakan untuk melatih model, sementara data uji, yang mencakup 20% sisanya, digunakan untuk mengevaluasi performa model yang telah dilatih.

3.4 Implementasi Model

Model dasar yang diterapkan adalah Random Forest dengan 100 estimators dan XGBoost Random Forest dengan 100 estimators. Selain itu, Logistic Regression digunakan sebagai meta-model untuk menghasilkan prediksi akhir.

3.5 Implementasi Stacking Classifier

implementasi Stacking Classifier dilakukan menggunakan pustaka Scikit-learn. Proses ini melibatkan beberapa langkah, yaitu mendefinisikan model dasar berupa Random Forest dan XGBoost, mendefinisikan meta-model Logistic Regression, menggabungkan semua model ke dalam Stacking Classifier, dan melatih model menggunakan data latih.

3.6 Evaluasi Model

Model ini dievaluasi menggunakan berbagai metrik seperti confusion matrix untuk melihat distribusi prediksi yang benar dan salah, akurasi untuk mengetahui proporsi prediksi yang benar, sensitivitas untuk mengevaluasi kemampuan model mendeteksi kasus positif, dan spesifisitas untuk mengevaluasi kemampuan model mendeteksi kasus negatif. Selain itu, nilai ROC AUC dihitung untuk menunjukkan performa model secara keseluruhan.

3.7 Visualisasi Metrik

visualisasi metrik dilakukan untuk mempermudah interpretasi hasil. Confusion matrix divisualisasikan dalam bentuk heatmap, sementara ROC curve digambarkan untuk menunjukkan keseimbangan antara sensitivitas dan spesifisitas model. Visualisasi ini membantu memberikan gambaran yang lebih jelas mengenai performa model yang dikembangkan.