

Nama : Ilham Novriadi

NIM : 231011403539

Bagian 3 – Analisis dan Kesimpulan

Analisis Hasil

Berdasarkan eksperimen yang telah dilakukan pada Bagian 2, berikut adalah analisis mendalam mengenai performa model Decision Tree:

1. **Model Terbaik:**
Model Decision Tree dengan parameter `max_depth=5` terbukti memberikan hasil yang paling seimbang.
 - **Tanpa pembatasan kedalaman**, pohon keputusan cenderung tumbuh sangat kompleks, menangkap setiap *noise* dalam data training. Hal ini menyebabkan akurasi training mencapai 100% namun akurasi testing menurun drastis (indikasi kuat *overfitting*).
 - **Dengan `max_depth=5`**, model dipaksa untuk melakukan generalisasi, sehingga performa pada data testing tetap tinggi dan stabil.
2. **Faktor yang Mempengaruhi Performa:**
 - **Preprocessing Data:** Langkah penghapusan data duplikat sangat krusial. Dataset `heart.csv` aslinya memiliki beberapa baris data yang identik. Tanpa menghapusnya, model bisa menjadi bias dan evaluasi akurasi menjadi tidak realistis karena kebocoran data (data yang sama muncul di train dan test).
 - **Hyperparameter Tuning:** Parameter `criterion` (menggunakan 'gini' atau 'entropy') memberikan hasil yang serupa, namun pengaturan `max_depth` adalah faktor penentu utama dalam mencegah *overfitting*.
 - **Fitur Signifikan:** Berdasarkan struktur pohon yang dihasilkan, fitur seperti `cp` (tipe nyeri dada), `thal` (thalassemia), dan `ca` (jumlah pembuluh darah utama) sering muncul di bagian atas pohon (root/internal nodes). Ini menunjukkan bahwa fitur-fitur medis ini adalah indikator terkuat untuk memprediksi penyakit jantung.

Kelebihan Tree-based Methods pada Studi Kasus Ini

Dalam konteks prediksi medis (Penyakit Jantung), metode berbasis pohon memiliki keunggulan spesifik:

- **Interpretabilitas Tinggi:** Tidak seperti *Neural Networks* yang sering dianggap sebagai "black box", Decision Tree menghasilkan aturan logika yang jelas (Contoh: "JIKA `cp > 0` DAN usia `< 60` MAKA..."). Hal ini sangat penting di dunia medis karena dokter perlu memahami *mengapa* sistem memprediksi seorang pasien sakit.
- **Penanganan Hubungan Non-Linear:** Faktor risiko kesehatan seringkali tidak berhubungan secara linear. Misalnya, kolesterol tinggi mungkin hanya berbahaya jika

disertai dengan usia tua atau tekanan darah tinggi. Decision Tree secara alami menangkap interaksi non-linear antar variabel ini melalui percabangan bertingkat.

- **Minimal Data Preprocessing:** Decision Tree tidak memerlukan penskalaan data (seperti standarisasi/normalisasi) sehingga menyederhanakan pipeline pemrosesan data, terutama karena dataset ini campuran antara data numerik dan kategorikal.

Kesimpulan Akhir

Penerapan algoritma **Decision Tree** pada dataset Heart Disease berhasil menghasilkan model klasifikasi dengan performa yang memadai untuk tujuan skrining awal.

- Model mampu membedakan pasien sehat dan berisiko penyakit jantung dengan akurasi yang dapat diterima setelah dilakukan tuning parameter.
- Meskipun Decision Tree tunggal sangat mudah dipahami, ia memiliki kelemahan dalam hal stabilitas (sedikit perubahan data bisa mengubah struktur pohon).
- **Rekomendasi:** Untuk pengembangan selanjutnya di lingkungan produksi medis yang membutuhkan akurasi dan reliabilitas lebih tinggi, disarankan untuk menggunakan metode *Ensemble* seperti **Random Forest** atau **Gradient Boosting (XGBoost/LightGBM)**. Metode ini akan menggabungkan kekuatan banyak pohon untuk mengurangi varian dan meningkatkan akurasi prediksi secara keseluruhan.