

1. Algoritma Decision Tree adalah model machine learning yang membangkitkan struktur pohon keputusan berdasarkan data training. Model ini bekerja dengan cara membagi data secara rekursif berdasarkan fitur dan threshold yang paling optimal, dengan tujuan memaksimalkan purity pada setiap pembagian. Decision Tree juga merupakan model yang powerful karena dianggap dapat menangani hubungan non-linear dalam data.

Cara kerja

- a. Inisialisasi Parameter
 - i. min_samples_split: Minimum sampel yang diperlukan untuk split node (default: 2)
 - ii. max_depth: Kedalaman maksimum pohon (default: 100)
 - iii. root: Node akar dari pohon keputusan
- b. Training (Fit method)
 - i. Menyimpan jumlah fitur
 - ii. Memulai proses pembangunan pohon dari root node
 - iii. Membangun pohon secara rekursif dengan metode `_grow_tree`
- c. Tree Construction (`_grow_tree`)
 - i. Memeriksa kondisi berhenti (stopping conditions):
 1. Kedalaman mencapai maksimum
 2. Hanya satu kelas yang tersisa
 3. Sample kurang dari min_samples_split
 - ii. Mencari split terbaik menggunakan `_best_split`
 - iii. Membangun left dan right subtree secara rekursif
- d. Mencari Split terbaik (`_best_split`)
 - i. Mengevaluasi semua fitur dan threshold yang mungkin
 - ii. Menghitung Gini Gain untuk setiap kemungkinan split
 - iii. Memilih split dengan Gini Gain tertinggi
- e. Gini Calculations
 - i. Gini Impurity: Mengukur kemurnian node ($1 - \sum p_i^2$)
 - ii. Gini Gain: Mengukur improvement setelah split ($\text{parent_gini} - \text{child_gini}$)
- f. Prediction (Predict method)
 - i. Traverse pohon untuk setiap sample test
 - ii. Pada setiap node, periksa kondisi fitur dan threshold
 - iii. Mengembalikan nilai leaf node ketika mencapai terminal

2. Berdasarkan hasil evaluasi, model dari Sklearn memiliki nilai Recall 0.60 dan F1 score 0.60, sementara untuk model yang saya buat sendiri mendapatkan nilai Recall 0.59 dan F1 score 0.59. Hal ini menunjukkan performa model yang sudah sangat mirip.
3. Beberapa improvement yang dapat dilakukan
 - a. Hyperparameter tuning dengan menggunakan tools seperti Grid Search, Optuna, dll untuk mendapatkan kombinasi parameter terbaik untuk model pada dataset ini.
 - b. Ensemble model yaitu menggabungkan beberapa model untuk melakukan prediksi sehingga kita dapat memperoleh model yang lebih robust dan mendapat akurasi yang lebih baik.
 - c. Menambahkan feature_importance untuk menghitung seberapa penting setiap fitur dengan mengakumulasi penurunan impurity setiap kali digunakan untuk split.