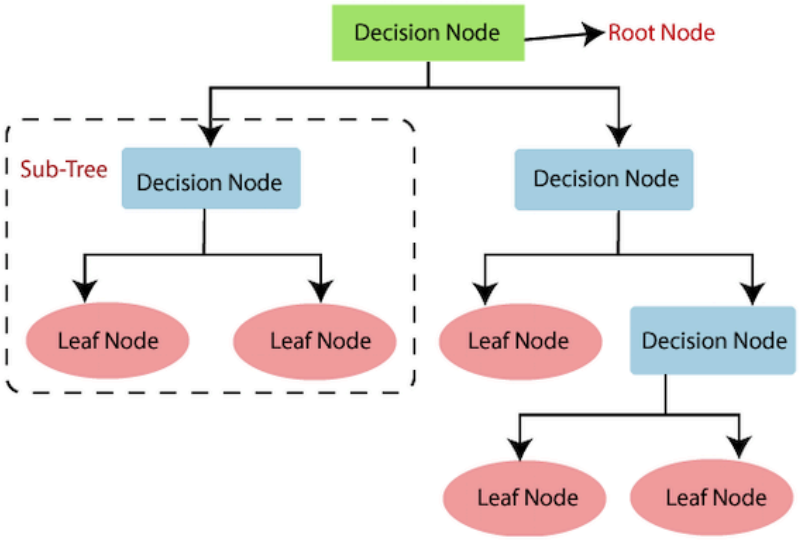
CART

Salman Hanif - 13523056

Cara Kerja Algoritma

Algoritma CART (Classification and Regression Trees) atau Decision Tree bekerja dengan cara memecah data menjadi subkelompok-subkelompok yang lebih homogen berdasarkan fitur-fitur yang ada. Cara pemecahan data adalah dengan syarat (memenuhi atau tidak), seperti pohon keputusan pada umumnya. Yang dilakukan oleh algoritma adalah pemilihan urutan fitur untuk seleksi dan pembuatan child nodes yang dilakukan secara rekursi sampai pohon terbentuk.



Langkah kerja untuk fungsi fit dan predict Decision Tree:

Proses pelatihan (fit):

1. Node root dibuat, pertama seluruh dataset masuk.
2. Dilakukan pengecekan apakah merupakan leaf node atau bukan dengan 3 kriteria stop : semua data di node labelnya sama, depth sudah max, jumlah data di bawah min data split
3. Jika masih bisa dipecah/belum leaf node, algoritma akan mencari fitur dan threshold untuk pemisahan di node tersebut dengan perhitungan Information Gain.
4. Pembagian data kepada child node, jika data kurang dari atau sama dengan threshold, masuk left child **(haluan kiri)**, kalau melebihi threshold masuk

right child.

Perhitungan information gain berdasarkan nilai entropy (keacakan data) dari parent dan child, InformationGain = Entropy(Data Parent) - Entropy(Data Child). Entropy data child ini rata-rata dari left child & right child.

Proses prediksi (predict):

Dilakukan penelusuran pada pohon dimulai dari root node, masuk ke setiap cabang untuk pembandingan nilai fitur dengan threshold, hingga mencapai leaf node. Modus label pada leaf node lah yang menjadi label bagi data baru.

Hasil Evaluasi model dari hasil scratch dan dari library. Jelaskan perbedaan

Dalam evaluasi model dengan hold-out dan k-fold, model yang saya buat menghasilkan akurasi dan f1-score yang lebih baik dari decision tree classifier scikit learn pada data yang saya gunakan. Model yang saya gunakan di-setting dengan minDataSplit 100 dan maxDepth 100. Kemungkinan ada potensi overfitting dengan maxDepth 100, tetapi harusnya bisa dicegah dengan minDataSplit 100 karena jumlah data 1800 an.

Setelah dicari tahu, pada perhitungan information gain terdapat perbedaan metode. Secara default, tree scikit-learn menggunakan Gini impurity, berbeda dengan saya yang menggunakan entropy. kemudian pruning saya menggunakan minDataSplit dan maxDepth, sedangkan parameter di fungsi scikit learn ada min_samples_leaf dan max_samples_leaf bukan depth.

Improvement yang bisa dilakukan

Improvement yang bisa saya lakukan adalah mencoba metode lain untuk pemisahan data pada decision tree seperti nilai Gini Impurity yang menggantikan entropy.