Random Forest Classifier

Deskripsi Algoritma

Random Forest adalah metode ensemble learning untuk klasifikasi dan regresi yang membentuk sekumpulan pohon keputusan (decision trees) pada sub-sampel dari dataset dan menggabungkan prediksi mereka untuk meningkatkan akurasi dan mengontrol overfitting. Pendekatan ini merupakan pengembangan dari algoritma **Decision Tree** yang rentan terhadap variansi tinggi, terutama ketika struktur pohon menjadi terlalu dalam atau terlalu bergantung pada outlier.

Random Forest mengatasi kelemahan tersebut dengan membangun banyak pohon pada sampel acak dari data pelatihan (bootstrap sampling) dan rata-rata prediksi mereka. Selain itu, pada setiap percabangan, algoritma hanya mempertimbangkan subset acak dari fitur, menambah elemen regularisasi melalui **bagging + feature randomness**.

Komponen dan Prosedur

1. **Bootstrap Aggregation (Bagging)**

Untuk setiap estimator (decision tree):

- Sebuah subset acak dari data pelatihan (dengan pengambilan sampel ulang) diambil sebagai data pelatihan khusus untuk pohon tersebut.
- Hal ini mengurangi korelasi antar pohon dan meningkatkan kemampuan generalisasi model.

2. **Random Subset of Features**

Alih-alih mempertimbangkan semua fitur pada setiap split, hanya subset acak yang dipertimbangkan. Nilai `max_features` mengatur ukuran subset ini, dengan opsi umum seperti:

- `"sqrt"`: akar dari jumlah total fitur.
- `"log2"`: log basis 2 dari jumlah fitur.
- `None`: semua fitur dipertimbangkan (menjadikan Random Forest mendekati Bagged Trees biasa).

3. **Prediksi Mayoritas**

Untuk klasifikasi, prediksi akhir dihasilkan melalui voting mayoritas dari prediksi seluruh pohon. Untuk probabilitas kelas, probabilitas dari masing-masing pohon dirata-ratakan.

4. **Kalkulasi Feature Importance**

Feature importance dihitung berdasarkan total penurunan impurity (misalnya Gini) yang dikaitkan dengan fitur tersebut di seluruh pohon. Skor ini kemudian dinormalisasi terhadap total penurunan impurity semua fitur.