Nama : Fadhil Dzikri Aqila

NIM : 1103213136

Kelas : TK-45-03

Machine Learning – Random Forest

Tree memiliki satu aspek yang menghalanginya untuk menjadi alat yang ideal untuk predictive learning, yaitu dalam akurasi. Tree hanya berfungsi baik dengan data yang digunakan untuk membuatnya. Saat ada data baru, tree menjadi tidak fleksibel dalam mengklasifikasikan sampel baru. Random Forests menggabungkan kesederhanaan decision tree dengan fleksibilitas yang menghasilkan peningkatan akurasi yang besar.

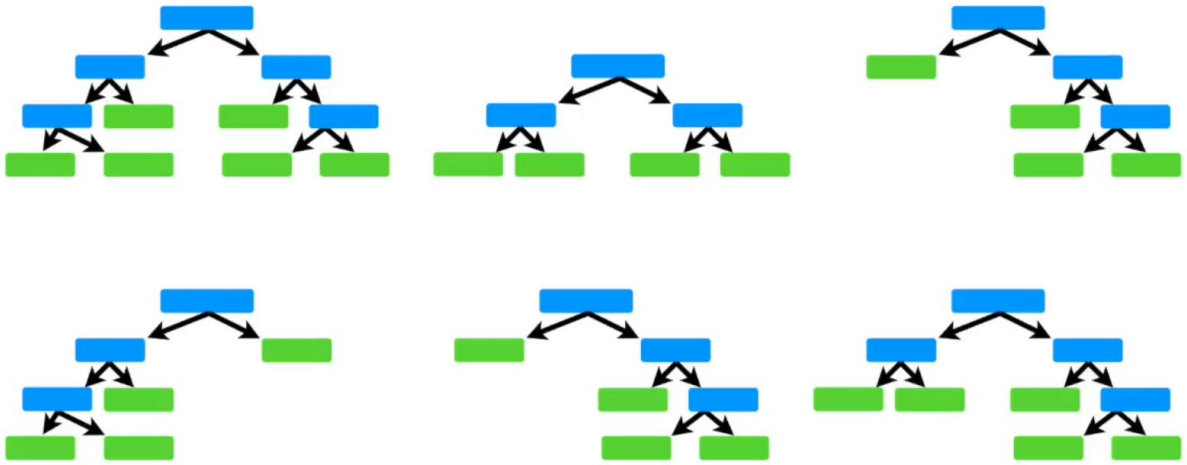
Berikut adalah langkah-langkah untuk membuat random forest :

1. Membuat kumpulan data bootstrap

   Untuk membuat kumpulan data bootstrap yang ukurannya sama dengan aslinya, kita cukup memilih sampel secara acak dari kumpulan data asli. Diperbolehkan untuk memilih sampel yang sama lebih dari satu kali.
2. Membuat decision tree menggunakan dataset bootstrap

   Hanya gunakan beberapa variabel atau kolom acak pada setiap langkah, misalkan terdapat 4 kolom, maka hanya dipilih 2 dari 4 kolom tersebut yang dijadikan kandidat untuk root. Karena salah satu node sudah terpilih, tersisa 3 kolom yang belum terpilih, kemudian dipilih lagi 2 dari 3 kolom yang dijadikan kandidat untuk node cabang. Kemudian lanjutkan dengan membuat tree seperti biasa, namun hanya mempertimbangkan subset variabel acak pada setiap langkah.
3. Kembali ke langkah 1 dan ulangi

   Buat kumpulan data bootstrap baru dan buat pohon dengan mempertimbangkan subkumpulan variabel di setiap langkah, idealnya harus dilakukan sebanyak ratusan kali, namun untuk contoh hanya dilakukan sebanyak 6 kali.



Gambar diatas adalah contoh hasil dari proses pembuatan random forests, dalam contoh hanya dibuat sebanyak 6 tree. Menggunakan sampel bootstrap dan hanya mempertimbangkan sebagian variabel pada setiap langkah akan menghasilkan beragam tree. Keanekaragaman inilah yang menjadikan random forests lebih efektif dibandingkan decision tree individual.

Cara penggunaan random forest dilakukan dengan menjalankan data baru melalui setiap decision tree yang telah dibuat sebelumnya. Setiap tree akan memberikan vote untuk kategori yang sesuai, dan hasil akhir diambil dari mayoritas suara.

Untuk mengetahui apakah hasil dari random forests bagus atau tidak, digunakan out-of-bag dataset. Biasanya sekitar 1/3 dari data asli tidak dimasukkan ke dalam kumpulan data bootstrap. Out-of-bag ini adalah data yang tidak termasuk dalam bootstrap dataset karena diperbolehkan adanya duplikasi. Data out-of-bag ini akan menjalankan seluruh random forests yang sebelumnya sudah dibuat, sehingga kita bisa mengukur akurasi model. Out-of-bag error dihitung sebagai proporsi sampel yang salah diklasifikasikan oleh random forest.

Penentuan jumlah variabel yang digunakan pada setiap step dalam pembuatan decision tree dapat dioptimalkan dengan mencoba beberapa pengaturan berbeda. Pengaturan yang memberikan random forest paling akurat akan dipilih setelah pengujian.