Nama : Eva Fiorina Siahaan
NIM : 1103210101

StatQuest: Random Forests Bagian 1 - Membangun, Menggunakan dan Mengevaluasi

Random Forest adalah jenis algoritma Machine Learning yang digunakan untuk memecahkan masalah regresi dan klasifikasi. Decision Tree mudah dibuat, mudah digunakan, dan mudah diinterpretasikan Decision Tree memiliki satu aspek yang menghalanginya untuk menjadi alat yang ideal untuk pembelajaran prediktif yakni dalam akurasi. Dengan kata lain, yang berfungsi dengan baik pada data yang digunakan untuk membuatnya tetapi tidak fleksibel dalam mengklasifikasikan sampel baru. Random forest menggabungkan kesederhanaan Decision Tree dengan fleksibilitas dengan menghasilkan peningkatan besar dalam akurasi.

Langkah - langkah :

- *Langkah 1*

membuat kumpulan data pada bootstrap yang ukurannya sama dengan aslinya, kemudian memilih sampel secara acak dari kumpulan data asli.

- *Langkah 2*

Membuat Decision Tree menggunakan dataset bootstrap namun hanya menggunakan sub kumpulan variabel atau kolom acak pada setiap langkah, contohnya hanya akan mempertimbangkan dua variabel atau kolom pada setiap langkah.

- *Langkah 3*

Mengulangi step 1 dan 2 beberapa kali, membuat data bootstrap baru. Menggunakan sampel bootstrap dan hanya akan mempertimbangkan sebagian variabel pada setiap langkah dan akan menghasilkan beragam pohon, inilah yang menjadikan random forest menjadi lebih efektif dibandingkan decision tree individual. Cara mengevaluasi random forest yaitu dengan menggunakan 2 variabel yang dimana kolom data untuk membuat keputusan pada tiap langkahnya. Lalu membandingkan error out-of-bag yang akan dibangun dengan menggunakan 2 variabel per langkahnya. Pertama dengan membangun random forest lalu yang kedua memperkirakan keakuratan random forest dan mengubah jumlah variabel yang akan digunakan pada tiap langkahnya kemudian melakukannya beberapa kali dan memilih yang paling akurat.

Salah satu cara agar Random Forest lebih optimal, ketika membuat pohon langkah pertama yang dilakukan hanya menggunakan dua variabel yaitu kolom data untuk membuat keputusan pada setiap langkah. Kemudian membandingkan error out-of-bag untuk Random Forest yang dibangun hanya dengan menggunakan dua variabel per-langkah ke Random Forest yang dibangun dengan menggunakan tiga variabel per langkah lalu menguji banyak pengaturan berbeda dan memilih Random Forest yang paling akurat.