K-Nearest Neighbors

Procedure 方法

K-Nearest Neighbors (KNN) adalah sebuah algoritma Machine Learning yang dapat digunakan untuk klasifikasi atau regresi. Caranya adalah dengan mencari K jumlah tetangga terdekat dan melakukan voting untuk klasifikasi dan mean untuk regresi.

1. **Inisialisasi**
   Model di-inisialisasi dengan parameter jumlah tetangga (K) dan metrik jarak yang diinginkan misalnya euclidean. Nilai K disini sangat memengaruhi hasil model, semakin kecil nilai K maka model akan lebih sensitif terhadap noise didalam tetangganya. Sebaliknya, semakin besar nilai K maka model dapat menjadi terlalu general.
2. **Menghitung jarak diantara data baru dengan semua baris data fit**
   Untuk setiap baris data baru, hitung jarak dengan semua baris data fit lainnya sesuai dengan metrik jarak yang dipilih sebelumnya.
3. **Pilih K buah baris data terdekat**
   Pada suatu baris data baru, K-buah baris data fit yang memiliki jarak paling kecil terhadapnya akan dipilih sebagai tetangga dari baris data tersebut.
4. **Voting untuk klasifikasi, mean untuk regresi**
   Dari target variable tetangga, lakukan voting untuk klasifikasi dan mean untuk regresi. Pada kasus draw untuk klasifikasi, pada model yang dibuat disini akan diambil value pertama yang muncul dengan scipy.stats.mode().

VS Sklearn

Hasil di notebook DoE menghasilkan nilai F1 yang sama. Namun, secara performance algoritma, KNN from scratch disini jauh lebih pelan daripada sklearn. Hal ini dikarenakan optimasi perhitungan distance yang dilakukan skelarn.

Potential Improvements

- Performance improvement pada perhitungan jarak
- Weighted KNN yang memprioritaskan tetangga yang lebih dekat