Rizqika Mulia Pratama
13522126

K-Nearest Neighbors

1. Cara kerja algoritma

Algoritma K-Nearest Neighbors (KNN) adalah salah satu metode machine learning yang digunakan untuk tugas klasifikasi dan regresi. Dalam konteks klasifikasi, KNN bekerja berdasarkan prinsip kesamaan atau kedekatan antara data yang ingin diklasifikasikan dengan data yang sudah ada.

1. Menentukan nilai K

   Nilai K dalam KNN adalah jumlah tetangga terdekat yang akan dipertimbangkan untuk menentukan kelas dari data baru. Pemilihan nilai K sangat penting karena dapat mempengaruhi akurasi model. Nilai K yang terlalu kecil bisa membuat model terlalu sensitif terhadap noise, sedangkan nilai K yang terlalu besar bisa membuat model terlalu umum.

2. Menghitung jarak

   Ketika ada data baru yang ingin diklasifikasikan, langkah pertama adalah menghitung jarak antara data baru tersebut dengan setiap data dalam dataset. Jarak yang paling umum digunakan adalah jarak Euclidean, tetapi bisa juga menggunakan jarak Manhattan atau jarak lainnya tergantung pada kebutuhan.

3. Menentukan tetangga terdekat

   Setelah menghitung jarak, langkah selanjutnya adalah mengurutkan semua data dalam dataset berdasarkan jarak terdekat ke data baru. Kemudian, pilih K data teratas yang memiliki jarak terdekat. Data-data ini disebut dengan tetangga terdekat.

4. Voting untuk klasifikasi

   Dari K terdekat yang telah dipilih, lakukan voting untuk menentukan kelas dari data baru. Kelas yang paling sering muncul di antara tetangga terdekat tersebut akan menjadi prediksi kelas untuk data baru. Misalnya, jika dari 5 tetangga terdekat, 3 diantaranya adalah kelas A dan 2 lainnya adalah kelas B, maka data baru akan diklasifikasikan sebagai kelas A.

5. Mengatasi ties

   Jika terjadi ties (misalnya, jumlah kelas yang sama dari tetangga terdekat), bisa digunakan beberapa strategi seperti memilih kelas dengan jarak rata-rata terdekat atau memilih secara acak.

6. Evaluasi

   Setelah klasifikasi dilakukan, model KNN dapat dievaluasi menggunakan data uji untuk mengukur akurasinya. Jika hasilnya tidak memuaskan, kita bisa menyesuaikan nilai K atau melakukan preprocessing data lebih lanjut seperti normalisasi atau penghapusan fitur yang kurang relevan.

Algoritma ini cukup efektif untuk dataset yang tidak terlalu besar.

Rizqika Mulia Pratama
13522126

2. Perbandingan model buatan sendiri dan library

Berdasarkan hasil evaluasi, kedua model memiliki nilai Recall dan F1-Score yang sama, yaitu 0.86 dan 0.83. Ini menunjukkan bahwa kedua model memiliki performa yang setara dalam hal mengukur kemampuan model untuk mendeteksi kelas positif (Recall) dan keseimbangan antara presisi dan recall (F1-Score)

3. Improvement yang bisa dilakukan

- Optimasi Hyperparameter

  Gunakan teknik seperti Grid Search atau Random Search untuk menemukan kombinasi parameter terbaik untuk model KNN, seperti n_neighbors, metric, dan p untuk metrik Minkowski

- Weighted KNN

  Mencoba gunakan versi berbobot dari KNN, di mana tetangga yang lebih dekat memiliki pengaruh lebih besar pada prediksi.

- Ensemble Methods

  Gunakan metode ensemble untuk menggabungkan prediksi dari beberapa model untuk meningkatkan akurasi