Laporan Tugas 2 Kecerdasan Buatan

K- Nearest Neighbor

M Reza Prawira S (1301161771)

1. Penyelesaian Masalah

1.1. Step step penyelesaian

1) Load data train awal dan data test dari csv. Pisahkan kolom atribut kelas(y) dengan atribut lain(X).
2) Prediksi kelas dengan parameter x train, y train, x validasi dan k.
    a. Split data train awal menjadi dua jenis data yaitu data train dan data validasi. Sehingga ada 4 variabel data yaitu X train,y train, X validasi dan y validasi. Data validasi berguna untuk mengetest hasil learning dari data train. Porsi pembagian adalah train > validasi. Porsi pembagian yang saya berikan adalah 27% data validasi dan sisanya data train.
    b. Hitung jarak antara X train dan X validasi dan simpan hasil perhitungan jarak ke list baru bernama distances.
    c. Urutkan list secara ascending dan cari label data dari k-data paling atas untuk dimasukkan ke list targets.
    d. Return target label paling banyak.
3) Lakukan prediksi untuk tiap data pada X test.
4) Hitung akurasi hasil prediksi dan cari akurasi terbaik dengan mengubah nilai k.
5) Jika sudah didapatkan akurasi paling baik, simpan nilai k terbaik dan lakukan prediksi pada data test asli (yang berjumlah 1000).
6) Simpan hasil prediksi pada Excel.

1.2.Analisis Masalah

Untuk menentukan nilai k, faktor penentu besarnya akurasi prediksi model adalah distribusi porsi data train. Dalam menentukan model, data yang sudah diketahui (data train keseluruhan) harus dibagi menjadi dua jenis data lagi yaitu data train dan data validasi. Data train berfungsi untuk membangun model sedangkan data validasi berfungsi sebagai cek performansi model yang telah dibangun.

Ada beberapa cara untuk membagi data menjadi data train dan validasi, pada percobaan ini saya melakukan pembagian data train dan data validasi secara random menggunakan modul dari library sklearn (train_test_split). Modul ini membagi porsi data train dan validasi secara random dengan besar proporsi besar data test yang dapat diatur oleh parameter test_size. Untuk mendapatkan model yang akurat, proporsi data validasi harus lebih kecil daripada data train. Hal ini dilakukan agar model mempunyai hasil yang lebih akurat karena melakukan learning data yang lebih banyak.

Berdasarkan hasil eksperimen, proporsi data train yang saya dapatkan adalah menggunakan data train sebanyak 73% dari data train asli karena menghasilkan akurasi yang lebih besar terhadap nilai k yang sama. Setelah mendapatkan proporsi data train yang sesuai, saya melakukan percobaan menentukan nilai k dengan range

antara 1 sampai 100. Percobaan saya lakukan antara 10 sampai 30 kali sehingga mendapatkan akurasi model terbesar yaitu 70.74 % untuk nilai k = 57.

```
E:\College Stuff\Materi\AI\Tugas2\Tugas Fix>python KNNsuci2.py
70.74074074074073 %
```

Setelah mendapatkan nilai k yang tepat maka saya menggunakan model untuk memprediksi data test asli dan menyimpan hasil prediksi pada excel berjudul 'Prediksi_Tugas2AI_[1301161771].csv'.

2. Kesimpulan

Model pembelajaran k-Nearest Neighbor merupakan model yang memprediksi target kelas data berdasarkan jarak antara data test dan data train. Untuk memperoleh nilai k terbaik perlu dilakukan penentuan besar proporsi data train asli dengan membagi data train asli menjadi data train dan data validasi. Jika sudah ditemukan besar proporsi data train asli yang sesuai maka eksprimen terhadap nilai k dapat dilakukan. Sehingga dapat disimpulkan parameter yang berpengaruh dalam observasi ini adalah besar proporsi data train dan nilai k.