NAMA = EDY SANTOSO NIM = 1301162111 KELAS=IF-40-08

DESKRIPSI MASALAH

Menentukan sebuah klasifikasi dari suatu class yang diberikan dengan beberapa atribut yang menjadi acuan untuk menentukan klasifikasinya. Laporan kali ini akan membahas problem dengan menggunakan methode K-Nearest Neighbors. Diberikan file DataTrain_Tugas3_AI.csv berupa himpunan data berisi 800 data yang memiliki 5 atribut input (X1, X2, X3, X4, dan X5) dan 1 output yang memiliki 4 kelas/label (0, 1, 2, dan 3). Problem yang diberikan adalah membangun sebuah sistem klasifikasi dengan methode K-NN untuk menetukan klasifikasi/label data testing dalam file DataTest_Tugas3.csv. Sistem akan membaca masukan file DataTrain_Tugas3_AI.csv dan DataTest_Tugas3_AI.csv dan mengeluarkan outpu berupa file TebakanTugas3.csv berupa satu kolom berisi 200 baris angka bernilai bulat/integer (0,1,2, atau 3) yang menyatakan kelas/label baris atau record yang bersesuaian pada file DataTest_Tugas3_AI.csv.

ANALISYS DAN STRATEGI

Strategi yang digunakan untuk menyelesaikan permasalahan diatas adalah membuat program denga methode K-NN dengan membaca file DataTest_Tugas3_AI.csv dan DataTrain_Tugas3_AI.csv dan mengeluarkan 200 data klasifikasi untuk file DataTest_Tugas3_AI.csv dalam satu kolom.

Dalam methode ini secara umum program akan membaca data train terlebih dahulu baru membaca data test untuk bisa membandingkan jarak antar setiap data test dengan seluruh data train. Setelah sistem selesai membandingkan maka akan dicari jarak terdekat sebanyak parameter K yang dipakai pada program K-NN.

Pada sistem K-NN normalnya akan membaca file data test. Pada setiap data test yang dibaca akan dicari jaraknya dengan data train. Setelah jarak didapatkan, maka jarak yang didapatkan akan diurutkan dan dicari data terdekat jaraknya sebanyak K. jadi jika dihitung kompleksitasnya maka didapatkan 200*(800+800*800)=128.160.000. Pada sistem K-NN yang dibangun ini tidak seperti pada umumnya. Program akan membaca File DataTest_Tugas3_AI.csv yang berisi 200 baris data baru membaca DataTrain. Setiap data train akan dibandingkan dengan dengan data test dan menyimpan data terdekat sebanyak K. setelah K terdapat nilai semuanya data train selanjutnya akan dibandingkan dengan nilai K tertinggi yang dimiliki oleh setiap data test. Jika data train memiliki jarak lebih dekat dibandingkan dengan jarak K terjauh dari data test maka data train akan disimpan untuk menggantikan data K terjaun. Begitu terus hingga data train dibaca keseluruhannya. Sehingga diperoleh kompleksitas yang lebih kecil, sebagai berikut perhitungannya 800*200+1=160.001.

Untuk menentukan nilai K yang paling optimal diperlukan data validasi yang diambil dari data train. Data train dibagi menjadi 2, bagian pertama 30% sebagai data validasi dan bagian kedua 70% sebagai data train.

Data pertama berisi 30%*800=240 data yang dijadikan sebagai data validasi terhadapa data train yang 70%. Data validasi berisi 80 baris pertama diambil dari data train no urut 1-80, 80 baris kedua diambil dari data train no urut 401-480, sedangkan 80 baris terakhir diambil dari data train no urut 721-800. Dan sisanya dijadikan sebagai data train untuk validasi, jadi terdapat 560 baris untuk data train validasi. Berikut data hasil pengukuran akurasi data validasi terhadap nilai K dari 1-20.

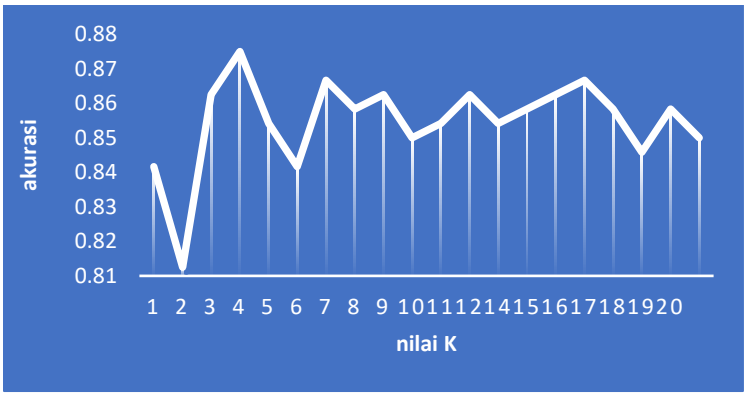

Figure 1. Hasil pengukuran

Dari hasil pengukuran data validasi terlihat bahwa K=4 memiliki akurasi yang cukup tertinggi yakni 87.5%. Berdasarakn data validasi yang didapatkan sehingga kami manarik kesimpulan untuk menggukan nilai K=4.

Berikut screenshot bukti hasil running terbaik dari program yang telah dibuat dengan parameter K=4.