Laporan Tugas 3 Pengantar Kecerdasan Buatan

I Kadek Andrean Pramana Putra
1301180179 / IF 42 08

1. Tujuan Utama

Diberikan dataset (himpunan data) Pima India Diabetes Dataset (PIDD) pada file "Diabetes.csv". Dataset tersebut berisi 768 objek data (baris). Buatlah lima datasets baru menggunakan skema 5-fold cross-validation. Pertama, bagi objek data ke dalam lima subsets (sub himpunan) dengan porsi yang sama,"

2. Spesifikasi Implementasi

Implementasi dari system yang telah dibangun mencakup

1. Pembangunan system menggunakan Bahasa python dengan menggunakan jupyter notebook sebagai platform utama untuk pengerjaan
2. Library yang dipakai dalam program:
   - Pandas agar bisa mengakses file Diabetes.CSV agar data dari file tersebut bisa di import datanya.
   - Numpy untuk mengatur agar data berbentuk matriks
   - MatPlotLib untuk memvisualisasikan dalam bentuk scatterplot
   - Collections untuk mereturn data yang keluar paling terbanyak setelah diproses
   - Sklearn.metrics. accuracy score yaitu untuk menghasilkan nilai akurasi dari data yang ada

3. Strategi penyelesaian

3.1 Perhitungan jarak

Metode yang saya gunakan dalam merancang program ini yaitu dengan metode manhattan dengan rumus,

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

Dimana *d* adalah jarak x merupakan atribut yang akan dilakukan perhitungan, sedangkan I dan j yaitu iterasi untuk tiap-tiap data yang tersedia dalam dataset.

3.2 Pre-processing Data

Dataset yang diberikan di preprocessing menggunakan rumus normalisasi Z score. Dengan formula ini, masing-masing nilai pada fitur dikurangi dengan miu (μ) yang merupakan nilai rata-rata fitur, kemudian dibagi dengan sigma (σ) yang merupakan standar deviasi.

$$x_{new} = \frac{x_{old} - \mu}{\sigma}$$

3.3 kNN

kolom dalam dataset pada program yang sudah dibuat dibedakan menjadi beberapa

atribut yaitu : **Pregnacies**, **Glukosa**, **BloodPressure**, **SkinThickness**, **Insulin**, **BMI**, **DiabetesPedigree**, **Age**

Label yaitu : **Outcome**

Atribut yang diekstrak dari Dataset, kemudian dilakukan preprocessing data menggunakan rumus normalisasi. Semua atribut dan label akan dibagi menjadi 5 bagian, meliputi ***Train***(data uji) dan ***Test***(data latih). Setiap ***Train*** (data uji) dan ***test*** (data latih) akan dihitung jaraknya menggunakan metode manhattan. Setelah terkumpul setiap jarak akan diurutkan dari yang terkecil. Dan akan diambil akurasi dari tiap nilai

Laporan Tugas 3 Pengantar Kecerdasan Buatan

I Kadek Andrean Pramana Putra
1301180179 / IF 42 08

3.4 5-Fold Cross Validation

Atribut dan label dibagi menjadi 5 bagian. Untuk tiap bagian memegang daa yang akan diklasifikasi dan setelah ***testing***(data uji) diberikan label ***train*** (data latih), akan dibandingkan dengan label ***testing*** (data uji) untuk mencari skor akurasinya yaitu antara (0-100). Maka tiap bagian data terselesaikan, akan muncul skor akurasi bagian tersebut. Setelah akurasi tiap lapis didapatkan, maka akan dilakukan penghitungan rata-rata dari setiap bagian nilai akurasi. Proses ini berulang terus menerus bergantung pada nilai k dari setiap skor akurasi rata-rata dari setiap bagian yang didapatkan, hasil akhir program akan memberikan nilai k terbaik berdasarkan skor rata-rata akurasi

4. Parameter

Pada program yang sudah dibuat, nilai maksimal k dibatasi menjadi 20 karena banyak menggunakan pengulangan dalam memproses data tersebut, maka dari itu nilai sekuens dari nilai k yaitu 1, .. 3, .. 5, .. 39. Untuk proses cross validation dilakukan berdasarkan setiap nilai k, grafik skor akurasi untuk tiap nilai k akan di tampilkan dalam bentuk grafik pada sub judul output system

5. Output system

```
Nilai K :  1
Akurasi :  70.33263510682866
Nilai K :  3
Akurasi :  71.8843736908253
Nilai K :  5
Akurasi :  72.40720569752827
Nilai K :  7
Akurasi :  74.7373271889401
Nilai K :  9
Akurasi :  75.64306661080855
Nilai K :  11
Akurasi :  74.99204021784666
Nilai K :  13
Akurasi :  74.47423544197737
Nilai K :  15
Akurasi :  75.38416422287389
Nilai K :  17
Akurasi :  75.1277754503561
Nilai K :  19
Akurasi :  76.42563887725177
Nilai K :  21
Akurasi :  76.55550900712191
Nilai K :  23
Akurasi :  76.55215751989945
Nilai K :  25
Akurasi :  75.90364474235443
Nilai K :  27
Akurasi :  75.90532048596565
Nilai K :  29
Akurasi :  76.03519061583577
Nilai K :  31
Akurasi :  75.90364474235443
Nilai K :  33
Akurasi :  76.42228739002933
Nilai K :  35
Akurasi :  76.42228739002933
Nilai K :  37
Akurasi :  76.81022203602849
Nilai K :  39
Akurasi :  76.16254713028907
Nilai k Terbaik : 37
Akurasi Tertinggi : 76.81022203602849
```