Tugas 3
CII-2M3 Pengantar Kecerdasan Buatan
Ganjil 2020/2021

Nama : Daniel Septyadi

Nim : 1301180009

Kelas : IF-42-07

1. Permasalahan

   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*. Bagi objek data ke dalam lima *subsets* (sub himpunan) dengan porsi yang sama, masing-masing berisi sat per lima (20%) data.

2. Analisis

```
if sort[k]==dist[m]:
    m = m+1
    if dist[m]=='0':
        nol = nol+1
    elif dist[m]=='1':
        satu = satu+1
```

Sebelum melalui proses vote terlebih dahulu, kita cek tetangga terdekat kalau tetangga terdekat labelnya 0, masukkan nol = nol +1 setelah itu di cek, mana label yang paling banyak , maka itu yang akan jadi label data terbaru.

Input nilai k = 7 dengan dataset = 3

```
(base) D:\COOLYEAH\DATABASE\PROJECT\S5\AI\TUPRO3>C:/Users/Asus/anaconda3/python.exe d:/COOLYEAH/DATABASE/PROJECT/S5/AI/TUPRO3/Tupro3_1301180009.py
Input nilai k: 7
Pilih sorting:
1.Baris ke-1 sampai baris ke-614 sebagai training set dan sisanya sebagai testing set
2.Baris ke-1 sampai baris ke-461 ditambah baris ke-615 sampai 768 sebagai training set dan yang lain sebagai testing set
3.Baris ke-1 sampai baris ke-307 ditambah baris ke-462 sampai 768 sebagai training set dan yang lain sebagai testing set
4.Baris ke-1 sampai baris ke-154 ditambah baris ke-308 sampai 768 sebagai training set dan yang lain sebagai testing set
5.Baris ke-155 sampai sampai 768 sebagai training set dan yang lain sebagai testing set

 Dataset yang dipilih adalah : 3
Data sudah di output sebagai 'tebakan.csv'
Total Prediksi yang benar:  115 / 154 data
Akurasinya adalah :  75%
```

Prediksi data input nilai k = 7 dan dataset = 3 adalah 115/154 data, lalu pada Akurasinya terdapat 75% , dapat diambil kesimpulan bahwa akurasi tertinggi terdapat pada input nilai k = 7 dan dataset yang dipilih 3. Menurut analisis saya k terbaik harusnya yg ganjil

karena labelnya itu genap jadi akan lebih valid mencari tetangganya menggunakan angka ganjil supaya sistemnya tidak bingung mau dibuat di label yang mana

3. Strategi Penyelesaian Masalah
   - **Import CSV**

     Hal pertama yang dilakukan adalah membaca file 'Diabetes.csv' yang telah ada di folder, dan di import ke python.
   - **Determine Input dan Output**
     - **a.** input nilai k
     - **b.** output :
       - Data sudah dioutputkan sebagai 'tebakan.csv'
       - Total prediksi
       - Akurasi
   - **Design the Fuctions**

```
def distance(x0,x1,x2,x3,x4,x5,x6,x7,x0t,x1t,x2t,x3t,x4t,x5t,x6t,x7t): #rumus jarak
    root = (x0t-x0)**2+(x1t-x1)**2 + (x2t-x2)**2 + (x3t-x3)**2 + (x4t-x4)**2 + (x5t-x5)**2 +(x6t-x6)**2+(x7t-x7)**2
    return math.sqrt(root)
```

     Menggunakan rumus euclidean distance untuk pengukuran jaraknya.

```
def voting(nol,satu):
    l_kategori=[]
    l_kategori.insert(0,nol)
    l_kategori.insert(1,satu)
    return (l_kategori)
```

     vote untuk klasifikasi, kalau label yang dimasukin ke 0 maka akan di insert ke nol , lalu kalau 1 maka akan diinsertkan ke satu
   - **Euclidean distance**

     Formula euclid merupakan salah satu **formula** yang digunakan untuk mengukur **jarak** dari 2 titik dengan menggunakan perhitungan matematis (metode heuristik).

     perhitungan jarak menggunakan rumus :

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

(p,**q)** = dua titik ruang-n Euclidean

(**q***i*,**p***i*) = vektor Euclidean, dimulai dari asal ruang (titik awal)

(*n*) = ruang n

- **Result Run Program**

```
Data sudah di output sebagai 'tebakan.csv'
Total Prediksi yang benar:  115 / 154 data
Akurasinya adalah :  75%
```