#### LAPORAN TUGAS 3 PENGANTAR KECERDASAN BUATAN

Nama: Mohammad Dwiantara Mahardhika

NIM: 1301184467

Kelas : IF-42-03

# 1. Deskripsi Masalah

Implementasi *k-nearest neigbor* (*kNN*) yang mencari nilai k terbaik untuk dataset (himpunan data) Pima India Diabetes Dataset (PIDD) pada file "Diabetes.csv" dengan skema validasi model 5-fold cross validation.

# 2. Strategi Penyelesaian Masalah

## 2.1. Spesifikasi

Program dibangun dengan Bahasa Python. Library yang digunakan antara lain csv untuk membaca file Diabetes.csv, math untuk menggunakan fungsi sqrt, dan pandas untuk visualisasi akurasi dalam scatter plot.

#### 2.2. Perhitungan Ukuran Jarak

Semua atribut pada dataset Diabetes.csv berbentuk atribut numerik sehingga perhitungan ukuran jarak menggunakan formula Euclidean Distance.

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{ip})^2}$$

### 2.3. Prapemrosesan Data

Dilakukan normalisasi terhadap dataset dengan menggunakan *min-max scaling*.

$$x_i' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Normalisasi ini menskalakan semua nilai pada dataset ke dalam interval [0,1].

# 2.4. Klasifikasi KNN

Klasifikasi dilakukan dengan mencari k data latih terdekat untuk setiap data uji. Kelas prediksi model ditentukan dari kelas yang paling banyak di antara k *neighbors* tersebut. Kemudian dilakukan pengecekan kelas actual dan kelas prediksi model.

#### 2.5. 5-fold cross-validation

Dataset yang diberikan dibagi menjadi 5 bagian yang masing-masing berisi satu per lima (20%) data. Setiap bagian atau *fold* akan menjadi data uji secara bergantian untuk dihitung akurasinya. Setelah itu dilakukan perhitungan rerata akurasi.

# 3. Fungsi dan Prosedur

## • Load csv()

Membaca file csv untuk dipindahkan ke dalam list.

### • Split(a, n)

Membagi list a menjadi n bagian.

#### • Euclidean distance(x1, x2)

Fungsi untuk melakukan perhitungan jarak numerik antara data x1 dengan x2.

# • Normalization(dataset)

Melakukan normalisasi dengan menggunakan *min-max scaling* untuk mengubah nilai dataset kedalam interval [0,1]

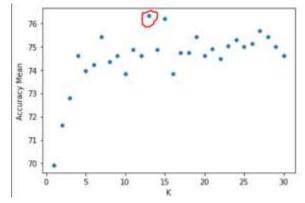
- **Get\_neighbors** (**train**, **test\_row**, **k**)
  Fungsi yang mengembalikan sejumlah *k* data latih pada *training* yang paling dekat dengan data uji *testing\_row*.
- Classification(train, test\_row, k)
  Fungsi untuk melakukan klasifikasi
  untuk satu data uji test\_row.
- K\_nearest\_neighbors(data\_test, data\_train, k)

  Melakukan klasifikasi untuk semua data yang ada pada data test.
- Accuracy(testing\_set\_outcome, prediction)

Melakukan perhitungan akurasi prediksi model terhadap nilai aktual data uji.

# 4. Hasil Observasi

Dilakukan observasi dengan mencari rerata akurasi untuk nilai k (k -nearest neighbors) = 1 sampai k = 30. Hasil rerata akurasi kemudian divisualisasikan dalam bentuk *Scatter Plot*.



Dari hasil observasi didapatkan rerata akurasi tertinggi didapatkan saat k = 13

yaitu 76,3%. Kemudian dilakukan pengecekan standar deviasi untuk memastikan bahwa k=13 adalah nilai k terbaik.

```
75.32467532467533, 79.08496732026144, 73.8562091503268]
Accuracy Mean: 74.87394957983193

K: 12
Accuracy List: [75.32467532467533, 69.48051948051948, 75.97402597402598, 78.43137254901961, 73.8562091503268]
Accuracy Mean: 74.61336049571344

EX: 13
Accuracy List: [76.62337662337663, 70.77922077922078, 76.62337662337663, 81.04575163398692, 76.47058823529412]
Accuracy Mean: 76.30846277905103

K: 14
Accuracy List: [76.62337662337663, 68.181818181817, 75.97402597402598, 79.08496732026144, 74.50980392156863]
Accuracy Mean: 74.87479840421017

EX: 15
Accuracy List: [76.62337662337663, 68.83116883116884, 76.62337662337663, 81.69934640522875, 77.12418300653596]
Accuracy Mean: 76.18029029793736
```

Dari hasil pengecekan standar deviasi nilai k = 13 mempunyai akurasi yang cukup stabil.

# 5. Output Program

Output program adalah rerata akurasi kNN untuk nilai k = 1 sampai k = 30 yang juga divisualisasikan dalam bentuk *Scatter plot* dan menampilkan nilai k terbaik beserta rerata akurasinya.

```
Accuracy Mean: 75.00721500721502

K: 26
Accuracy Mean: 75.13623631270691

K: 27
Accuracy Mean: 75.65741448094388

K: 28
Accuracy Mean: 75.39767422120363

K: 29
Accuracy Mean: 75.00721500721502

K: 30
Accuracy Mean: 74.61845344198287

Best K Value: 13
Accuracy Mean: 76.30846277905103

In [11]:
```