

# LAPORAN TUGAS 3 PENGANTAR KECERDASAN BUATAN

Nama : Mohammad Dwiantara Mahardhika

NIM : 1301184467

Kelas : IF-42-03

## 1. Deskripsi Masalah

Implementasi *k-nearest neighbor* (*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*.

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

### 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 akurasi. 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.

