PT UNIVERSAL BIG DATA
Ruko Modern Kav A16-A17, Jl Loncat Indah, Tasikmadu, Kota Malang 65143
No. Telepon 0812-1212-2388, Email : suratkita@gmail.com

Latihan Soal LKS AI UBIG

Prediksi Kekambuhan Kanker Tiroid (LEVEL SUPRA HARDCORE 💀🔥)

- **Dataset: filtered_thyroid_data**.csv

  Dataset ini berisi 383 data pasien kanker tiroid pasca terapi Radioactive Iodine (RAI). Setiap data mencakup informasi klinis, klasifikasi kanker, dan status kekambuhan. Tujuannya adalah memprediksi apakah kanker akan kambuh kembali (Recurred) berdasarkan data yang tersedia.

- **Tujuan**

  Membuat model **KNN klasifikasi manual (tanpa scikit-learn)** untuk memprediksi kemungkinan **kekambuhan kanker (Recurred)** berdasarkan data klinis pasien.

- **Spesifikasi Dataset**
  - Jumlah data: 383 pasien
  - Jumlah fitur: 13 kolom
  - Target klasifikasi: Recurred → Yes atau No
  - Tipe data: Mayoritas kategorikal
- **Kolom dalam Dataset**
  1. **Age**: usia pasien (numerik)
  2. **Gender**: jenis kelamin (M, F)
  3. **Hx Radiothreapy**: riwayat radioterapi (Yes, No)
  4. **Adenopathy**: pembesaran kelenjar getah bening (Yes, No)
  5. **Pathology**: tipe kanker (contoh: Micropapillary)
  6. **Focality**: jumlah fokus tumor (Uni-Focal, Multi-Focal)
  7. **Risk**: klasifikasi risiko kanker (Low, Intermediate, High)
  8. **T**: klasifikasi tumor (T1a, T2, dsb)
  9. **N**: klasifikasi nodus limfa (N0, N1)
  10. **M**: metastasis (M0, M1)
  11. **Stage**: stadium kanker (I, II, III, IV)
  12. **Response**: hasil respon pengobatan (Excellent, Indeterminate, dll)
  13. **Recurred**: kekambuhan (Yes, No) ← target klasifikasi
- **Tahap 1: EDA**
  1. Hitung distribusi jumlah pasien berdasarkan:
     1. Gender
     2. Risk level
     3. Response terhadap terapi
     4. Recurred
  2. Visualisasikan hubungan antara usia dan Recurred (boxplot).

3. Buat stacked bar chart: Stage vs Recurred.
4. Deteksi outlier pada usia menggunakan IQR dan Z-score.

- **Tahap 2: Data Pre-processing**
  1. Konversi semua kolom kategorikal menjadi numerik secara manual (bukan one-hot dari library).
  2. Normalisasi Age (min-max scaling manual).
  3. Split data menjadi 80% training dan 20% testing tanpa train_test_split.
  4. Buat 2 fitur baru (opsional tapi bernilai tambah):
     - Risk_Score: Low = 0, Intermediate = 1, High = 2
     - Spread_Level: nilai gabungan dari T + N + M
- **Tahap 3: Problem Solving – Implementasi KNN Manual (No Library ML!)**
  1. Buat fungsi:
     - Jarak Euclidean & Manhattan manual
     - Cari tetangga terdekat (k-nearest)
     - Voting mayoritas dari tetangga untuk prediksi
  2. Uji model dengan beberapa nilai k (3, 5, 7)
  3. Bandingkan hasil prediksi menggunakan metode jarak berbeda
- **Tahap 4: Evaluasi Model**
  1. Buat confusion matrix manual
  2. Hitung:
     - Accuracy
     - Precision
     - Recall
     - F1-score
  3. Analisis:
     - Apakah model terlalu banyak false negative?
     - Bagaimana hasil prediksi untuk pasien risiko tinggi?

🎯 Final Boss: Challenge

Prediksikan status pasien berikut menggunakan modelmu (k = 5, Euclidean):

PT UNIVERSAL BIG DATA
Ruko Modern Kav A16-A17, Jl Loncat Indah, Tasikmadu, Kota Malang 65143
No. Telepon 0812-1212-2388, Email : suratkita@gmail.com

Age = 58

Gender = F

Hx Radiothreapy = Yes

Adenopathy = Yes

Pathology = Papillary

Focality = Multi-Focal

Risk = High

T = T3

N = N1

M = M0

Stage = Stage III

Response = Indeterminate

Apakah pasien ini akan mengalami kekambuhan (Recurred = Yes / No)?

Library yang diperbolehkan: **numpy, pandas, matplotlib, seaborn**

Library yang dilarang: **Semua library ML seperti xgboost, tensorflow, lightgbm, sklearn.**

🔥 LEVEL SUPRA HARDCORE – Semua serba manual. Logika harus tajam. Tidak ada bantuan model siap pakai!