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

Klasifikasi Buah (LEVEL HARDCORE 💀🔥)

- **Dataset**: fruit_data_with_colors.txt
- **Tujuan**: Peserta harus membangun model K-Nearest Neighbors (KNN) secara manual tanpa menggunakan scikit-learn. Model ini akan digunakan untuk memprediksi jenis buah berdasarkan berat, ukuran, dan skor warna.
- **Tantangan**:
  - Peserta tidak boleh menggunakan library sklearn
  - Semua perhitungan jarak, normalisasi, klasifikasi, dan evaluasi model harus dibuat dari nol!
  - Eksperimen mendalam dengan berbagai metrik jarak & pemilihan K terbaik secara manual!
- **Kolom dalam Dataset**
  1. fruit_label → Label numerik untuk jenis buah (1 = Apple, 2 = Mandarin, 3 = Orange, 4 = Lemon).
  2. fruit_name → Nama buah (apple, mandarin, orange, lemon).
  3. fruit_subtype → Subtipe dari buah (granny_smith, braeburn, turkey_navel, dll).
  4. mass → Berat buah dalam gram.
  5. width → Lebar buah dalam cm.
  6. height → Tinggi buah dalam cm.
  7. color_score → Skor warna buah (mungkin terkait dengan tingkat kematangan atau kesegaran).
- **Tahap 1: EDA**
  1. Hitung jumlah setiap jenis buah dalam dataset (visualisasi bar chart atau pie chart).
  2. Visualisasikan hubungan antara mass, width, dan height dalam scatter plot 3D.
  3. Buat histogram untuk melihat distribusi color_score per jenis buah.
  4. Gunakan heatmap untuk melihat korelasi antar fitur.
  5. Deteksi outlier menggunakan IQR dan Z-score pada mass dan color_score.
  6. Analisis apakah ada perbedaan signifikan antara "subtype" dalam satu jenis buah (gunakan boxplot atau ANOVA).

  > 💡
  > - **Apakah ada buah yang memiliki skor warna yang aneh dibanding yang lain? 🍏🍊🍋**
  > - **Jangan hanya lihat angka, tapi juga hubungan antar fitur!**

- **Tahap 2: Data Pre-processing**

  Menyiapkan dataset agar siap digunakan dalam model KNN buatan sendiri.

  1. Konversi fruit_name menjadi label numerik untuk klasifikasi.

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

2. Ubah fruit_subtype menjadi fitur numerik menggunakan One-Hot Encoding secara manual.
3. Buat fitur baru density = mass / (width * height) untuk melihat kepadatan buah.
4. Lakukan normalisasi fitur numerik (mass, width, height, color_score) menggunakan rumus normalisasi manual.
5. Pisahkan dataset menjadi 80% training dan 20% testing secara manual (tanpa train_test_split).
6. Gunakan PCA (Principal Component Analysis) secara manual untuk mengurangi dimensi fitur dan analisis apakah performa meningkat atau tidak.
7. Lakukan Feature Selection secara manual untuk memilih fitur yang paling berpengaruh terhadap klasifikasi.

> - **Bagaimana cara menghitung normalisasi dan PCA tanpa sklearn?** 🤯
> - **Apakah fitur "density" lebih baik dari "mass" dalam membedakan buah?**

- **Tahap 3: Implementasi Machine Learning (KNN)**

  Membangun dan mengoptimalkan model klasifikasi menggunakan KNN.

  1. Bangun algoritma KNN secara manual tanpa sklearn!
  2. Hitung jarak antara data uji dan data latih secara manual menggunakan:
     1. Euclidean Distance
     2. Manhattan Distance
     3. Minkowski Distance
  3. Tentukan tetangga terdekat (K-terdekat) secara manual tanpa fungsi bawaan.
  4. Gunakan metode Grid Search manual untuk mencari K terbaik.
  5. Bandingkan hasil klasifikasi menggunakan nilai K yang berbeda.
  6. Prediksi jenis buah untuk data berikut menggunakan KNN buatan sendiri:
     1. **Mass:** 160g
     2. **Width:** 7.2 cm
     3. **Height:** 7.4 cm
     4. **Color Score:** 0.80

> - **Peserta harus menghitung jarak antara titik data sendiri, tanpa sklearn!** 😱
> - **Bagaimana cara menemukan tetangga terdekat secara manual?**
> - **!**

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

- **Tahap 4: Evaluasi Model**

  Menilai performa model KNN dan melakukan optimasi lebih lanjut.

  1. **Evaluasi model dengan:**
     1. **Confusion Matrix (dibuat manual, tanpa sklearn!)**
     2. **Precision, Recall, dan F1-Score (hitung manual, tanpa sklearn!)**
  2. Analisis False Positives & False Negatives dalam Confusion Matrix.
  3. Coba optimasi model dengan balancing data atau menghapus fitur yang kurang relevan.
  4. Simpulkan apakah model ini cukup akurat untuk digunakan dalam klasifikasi buah di industri pang

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

Library yang dilarang: **sklearn, scipy**