



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 Penyakit Jantung

• Tahap 1: EDA

Peserta diminta untuk **menganalisis dataset heart_disease.csv** dan menjawab pertanyaan berikut:

1. Berapa jumlah pasien yang memiliki penyakit jantung dan tidak? (Gunakan visualisasi pie chart atau bar chart).
2. Buat histogram atau boxplot untuk melihat distribusi usia pasien dan chol (kolesterol).
3. Apakah ada korelasi antara thalach (detak jantung maksimum) dan target? Gunakan scatter plot dan hitung korelasinya!
4. Gunakan heatmap untuk menampilkan korelasi antar fitur. Sebutkan 3 fitur yang paling berhubungan dengan target!
5. Apakah ada outlier dalam chol dan trestbps? Gunakan IQR method untuk mendeteksinya.



Jangan hanya melihat rata-rata! Beberapa fitur lebih informatif kalau dibandingkan antara **pasien dengan penyakit jantung dan tanpa penyakit jantung!** 🤔

• Tahap 2: Data Pre-processing

Peserta harus **menyiapkan dataset sebelum digunakan dalam Machine Learning**.

1. Ubah fitur kategorikal seperti sex, cp, restecg, slope, dan thal menjadi numerik dengan One-Hot Encoding.
2. Normalisasi fitur numerik (age, chol, thalach, oldpeak, dll) menggunakan StandardScaler.
3. Lakukan feature selection untuk memilih fitur yang paling berpengaruh terhadap prediksi.
4. Pisahkan dataset menjadi 80% training set dan 20% testing set.
5. Periksa apakah dataset imbalance? Jika iya, bagaimana cara menyeimbangkannya?



Tidak semua fitur penting, ada yang bisa dihapus! Cari tahu mana fitur yang punya korelasi tinggi! 🚀

• Tahap 3: Implementasi Machine Learning (KNN)

Peserta akan mengimplementasikan **K-Nearest Neighbors (KNN)** untuk klasifikasi dalam memprediksi penyakit jantung.

1. Bangun model KNN Classifier untuk memprediksi apakah seseorang memiliki penyakit jantung.
2. Coba beberapa nilai K dan gunakan Elbow Method untuk menemukan nilai K terbaik.
3. Evaluasi model dengan Confusion Matrix, Precision, Recall, dan F1-Score.
4. Bandingkan model KNN dengan Logistic Regression! Model mana yang lebih baik?



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

5. Buat prediksi untuk pasien baru dengan data berikut:

1. Usia: 55 tahun
2. Jenis Kelamin: Laki-laki
3. Nyeri Dada: 2
4. Tekanan Darah: 140 mm Hg
5. Kolesterol: 220 mg/dl
6. Gula Darah Puasa: 0
7. Detak Jantung Maksimum: 165
8. Angin akibat olahraga: 0
9. Oldpeak: 1.5
10. Slope: 1
11. CA: 1
12. Thal: 2



Pilih nilai K dengan bijak! Jangan sampai model jadi overfitting atau underfitting! 🤔

- **Tahap 4: Evaluasi Model**

Peserta harus **menganalisis performa model dan mencari cara untuk meningkatkannya.**

1. Analisis False Positives dan False Negatives dalam Confusion Matrix. Mana yang lebih berbahaya dalam kasus penyakit jantung?
2. Coba optimasi model dengan mengubah nilai K atau menggunakan metode lain (misal, Decision Tree).
3. Lakukan cross-validation untuk mendapatkan estimasi akurasi yang lebih stabil.
4. Kesimpulan: Apakah KNN cukup baik untuk diagnosis penyakit jantung? Jika tidak, apa alternatifnya?

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