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 Kelangsungan Hidup Titanic (LEVEL HARDCORE 💀🔥)

- **Dataset**: train.csv
- **Tujuan**: Memprediksi apakah seorang penumpang Titanic selamat atau tidak.
- **Tantangan**: Soal ini 100% sulit, butuh kombinasi statistik, data engineering, dan optimasi model!

- **Tahap 1: EDA**

  Analisis mendalam terhadap dataset untuk memahami pola dan anomali.

  1. **Cari tahu apakah ada bias gender dalam kelangsungan hidup penumpang.** (Gunakan visualisasi **stacked bar chart**.)
  2. **Analisis distribusi usia penumpang yang selamat dan tidak selamat.** (Gunakan **KDE plot** dan hitung **mean & median per kelompok**.)
  3. **Buktikan apakah kelas tiket (Pclass) dan harga tiket (Fare) berpengaruh terhadap keselamatan.**
  4. Gunakan **boxplot** untuk melihat distribusi harga tiket per kelas.
  5. Gunakan **korelasi Pearson** antara Pclass, Fare, dan Survived.
  6. **Analisis hubungan antara jumlah keluarga (SibSp + Parch) dan tingkat keselamatan.**
  7. **Gunakan metode IQR & Z-score untuk mendeteksi dan menghapus outlier di Fare dan Age.**
  8. **Gunakan heatmap untuk menemukan fitur yang paling berkorelasi dengan Survived.**

  > 💡
  > - **Ada fitur yang tampaknya tidak berguna, tapi sebenarnya sangat penting. Jangan langsung dihapus! 😏**
  > - **Banyak anomali pada data! Apakah penumpang yang membayar lebih mahal lebih mungkin selamat? 🚤**

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

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

  1. **Tangani nilai yang hilang:**
     1. Age → Gunakan **KNN Imputer** atau regresi berdasarkan fitur lain.
     2. Cabin → Jangan dihapus! Buat fitur baru **"HasCabin" (0 = Tidak, 1 = Ya)**.
     3. Embarked → Isi dengan modus.
  2. **Pisahkan "Title" dari Name dan gunakan sebagai fitur baru.** (Mr, Mrs, Miss, Master, dll.)
  3. **Ubah Sex, Embarked, Title menjadi numerik menggunakan One-Hot Encoding.**
  4. **Buat fitur baru FamilySize = SibSp + Parch + 1.**
  5. **Konversi Ticket menjadi angka berdasarkan frekuensi kemunculannya.**

6. **Normalisasi Fare, Age, dan FamilySize menggunakan StandardScaler.**
7. **Lakukan Feature Selection untuk memilih fitur yang benar-benar relevan.**

> - **"Title" sangat berpengaruh! Apakah "Master" lebih sering selamat dibanding "Mr"? 👀**
> - **"Cabin" seolah tidak berguna, tapi apakah orang dengan kabin lebih eksklusif lebih mungkin selamat?**

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

Membangun model prediksi kelangsungan hidup dengan berbagai teknik.

1. Bangun model KNN Classifier untuk memprediksi Survived.
2. Gunakan GridSearchCV untuk menemukan nilai K, weights, dan metric terbaik.
3. Bandingkan model KNN dengan model lain:
    1. **Logistic Regression**
    2. **Random Forest**
    3. **XGBoost**
    4. **SVM**
4. Gunakan metode feature importance dari Random Forest/XGBoost untuk melihat fitur paling berpengaruh.
5. Coba ensemble model (voting classifier) untuk meningkatkan akurasi.
6. Prediksi apakah seorang penumpang berikut akan selamat atau tidak:
    1. **Title:** Miss
    2. **Jenis Kelamin:** Perempuan
    3. **Kelas Tiket:** 3
    4. **Usia:** 21 tahun
    5. **Jumlah Keluarga:** 0
    6. **Harga Tiket:** 8.5
    7. **Embarked:** S

> - **KNN bisa jadi buruk untuk dataset ini! Bisakah kamu membuktikan model lain lebih baik? 🤔**
> - **Optimasi parameter bisa meningkatkan skor F1 hingga lebih dari 85%!**

- **Tahap 4: Evaluasi Model**

Menilai performa model dan melakukan optimasi.

1. Evaluasi model dengan:
    1. **Confusion Matrix**

2. **Precision, Recall, F1-Score**
3. **ROC-AUC Score**

2. Analisis False Positives & False Negatives dalam Confusion Matrix.
3. Gunakan SHAP atau Permutation Importance untuk memahami bagaimana setiap fitur memengaruhi prediksi.
4. Jika akurasi model masih di bawah 85%, lakukan optimasi lebih lanjut.
5. Simpulkan apakah model ini cukup baik untuk dipakai dalam dunia nyata.

> 💡
> - **False Negative bisa berbahaya: Jika model salah memprediksi orang yang bisa selamat sebagai tidak selamat, dampaknya besar!** 😨
> - **Akurasi tinggi bukan segalanya! Model harus bisa diinterpretasikan!**

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