Nama : Muhar Ferdiansyah
NIM : 231011401057
Kelas : 05TPLE005
Mata Kuliah : Machine Learning
Dosen : AGUNG PERDANANTO S.Kom, M.Kom

---

1. Deskripsi Dataset

Dataset yang digunakan adalah **Pima Indians Diabetes Database** yang berisi data medis pasien dengan 8 fitur utama:

Fitur	**Deskripsi**
Pregnancies	Jumlah kehamilan
Glucose	Konsentrasi glukosa plasma
BloodPressure	Tekanan darah diastolik (mm Hg)
SkinThickness	Ketebalan kulit trisep (mm)
Insulin	Insulin serum 2 jam (mu U/ml)
BMI	Body Mass Index
DiabetesPedigreeFunction	Fungsi silsilah diabetes
Age	Usia (tahun)

Target:

- 0 = Tidak diabetes
- 1 = Diabetes

Karakteristik:

- Data tidak seimbang (imbalanced)
- Korelasi antar fitur medis
- Potensi missing value/outlier

2. Model yang Digunakan

Lima algoritma klasifikasi yang digunakan:

- **Logistic Regression**
- **Decision Tree**
- **K-Nearest Neighbors (KNN)**
- **Support Vector Machine (SVM)** (dengan hyperparameter tuning)
- **Random Forest** (dengan hyperparameter tuning)

Preprocessing:

- Imputasi missing value dengan median
- Scaling fitur numerik (StandardScaler)
- Split data: 80% train, 20% test

3. Hasil Evaluasi dan Pembahasan

Tabel Perbandingan Hasil

Model	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.81	0.78	0.74	0.76	0.86
Decision Tree	0.75	0.70	0.72	0.71	0.77
K-Nearest Neighbors	0.78	0.74	0.72	0.73	0.82
Support Vector Machine	0.83	0.80	0.76	0.78	0.88
Random Forest	0.85	0.82	0.80	0.81	0.90

ROC Curve

- **AUC tertinggi**: Random Forest (0.90), diikuti SVM (0.88)
- **Recall tinggi** penting untuk mengurangi kasus diabetes yang tidak terdeteksi (false negative)
- **Precision tinggi** mengurangi diagnosa salah (false positive)

Insight

- Fitur Glucose, BMI, dan Age sangat berpengaruh pada prediksi
- Model ensemble (Random Forest) memberikan performa paling stabil
- Evaluasi menggunakan banyak metrik memberikan gambaran komprehensif

Kesimpulan

Semua model yang digunakan menunjukkan performa yang bervariasi, mencerminkan kompleksitas data medis pada kasus prediksi diabetes. Hyperparameter tuning terbukti efektif dalam meningkatkan performa model SVM dan Random Forest, di mana model ensemble seperti Random Forest cenderung memberikan hasil yang lebih stabil. Dalam konteks medis, recall menjadi prioritas utama untuk meminimalkan risiko missed diagnosis, sehingga pasien yang benar-benar menderita diabetes dapat terdeteksi dengan baik. Evaluasi menggunakan berbagai metrik seperti akurasi, precision, recall, F1-score, dan AUC memberikan pemahaman yang lebih komprehensif terhadap performa model. Namun, implementasi model dalam dunia nyata tetap memerlukan validasi klinis lebih lanjut serta pertimbangan etika, terutama terkait interpretabilitas, bias, dan keamanan data pasien.