**🔹 a. Deskripsikan apa masalah yang sedang dicari solusinya**

Masalah yang ingin diselesaikan adalah **memprediksi apakah seseorang mengidap diabetes atau tidak** berdasarkan data hasil pemeriksaan medis seperti kadar glukosa, tekanan darah, insulin, dan lainnya. Tujuannya adalah membantu dalam **diagnosis awal penyakit diabetes** menggunakan model pembelajaran mesin, agar penanganan medis bisa dilakukan lebih cepat dan akurat.

**🔹 b. Apa jenis masalahnya (klasifikasi/regresi)? Dan apa isi datanya?**

* ✅ **Jenis Masalah**: **Klasifikasi**
  + Karena target output (kolom Outcome) berupa **dua kelas**:
    - 0: Tidak mengidap diabetes
    - 1: Mengidap diabetes
* 📊 **Isi Datanya** (dari dataset *Pima Indians Diabetes*):
  + Pregnancies (Jumlah kehamilan)
  + Glucose (Kadar glukosa dalam darah)
  + BloodPressure (Tekanan darah)
  + SkinThickness (Ketebalan kulit lipatan triceps)
  + Insulin (Kadar insulin serum)
  + BMI (Indeks Massa Tubuh)
  + DiabetesPedigreeFunction (Riwayat keluarga terhadap diabetes)
  + Age (Usia)
  + Outcome (Label: 1 untuk diabetes, 0 untuk tidak)

**🔹 c. Ceritakan tahapan menyelesaikan masalah ini**

Berikut tahapan penyelesaiannya:

1. **Persiapan Data**
   * Load dataset diabetes.csv
   * Cek & ganti nilai nol pada fitur medis (Glucose, BloodPressure, BMI, dll) dengan **nilai median** (karena nol bukan nilai valid)
2. **Split Data**
   * Pisahkan data menjadi training dan testing (80:20)
3. **Pembangunan Model**
   * Gunakan DecisionTreeClassifier dari Scikit-learn
4. **Evaluasi**
   * Gunakan data testing untuk evaluasi akurasi dan metrik lain
5. **Visualisasi**
   * Gambar pohon keputusan menggunakan plot\_tree() dari Scikit-learn

🧭 **(opsional skema)**:

**🔹 d. Apa Error metric yang anda gunakan? Jelaskan mengapa memilih itu**

**Metrik yang digunakan:**

* **Accuracy** → Persentase prediksi yang benar
* **Confusion Matrix** → Untuk melihat detail prediksi benar dan salah
* **Classification Report** → Menyajikan Precision, Recall, F1-score

**Alasan:**

* Karena ini masalah klasifikasi biner, metrik seperti **Precision dan Recall penting**, terutama jika kasus diabetes memiliki **konsekuensi tinggi** saat salah prediksi (false negative).
* Accuracy saja tidak cukup jika data tidak seimbang, maka kami sertakan juga F1-score dan confusion matrix.

**🔹 e. Bagaimana kinerja model anda? Berikan evaluasi dan rekomendasinya**

📈 **Evaluasi Model:**

* **Akurasi**: 72.73%
* **Confusion Matrix**:

lua

SalinEdit

[[74 25]

[17 38]]

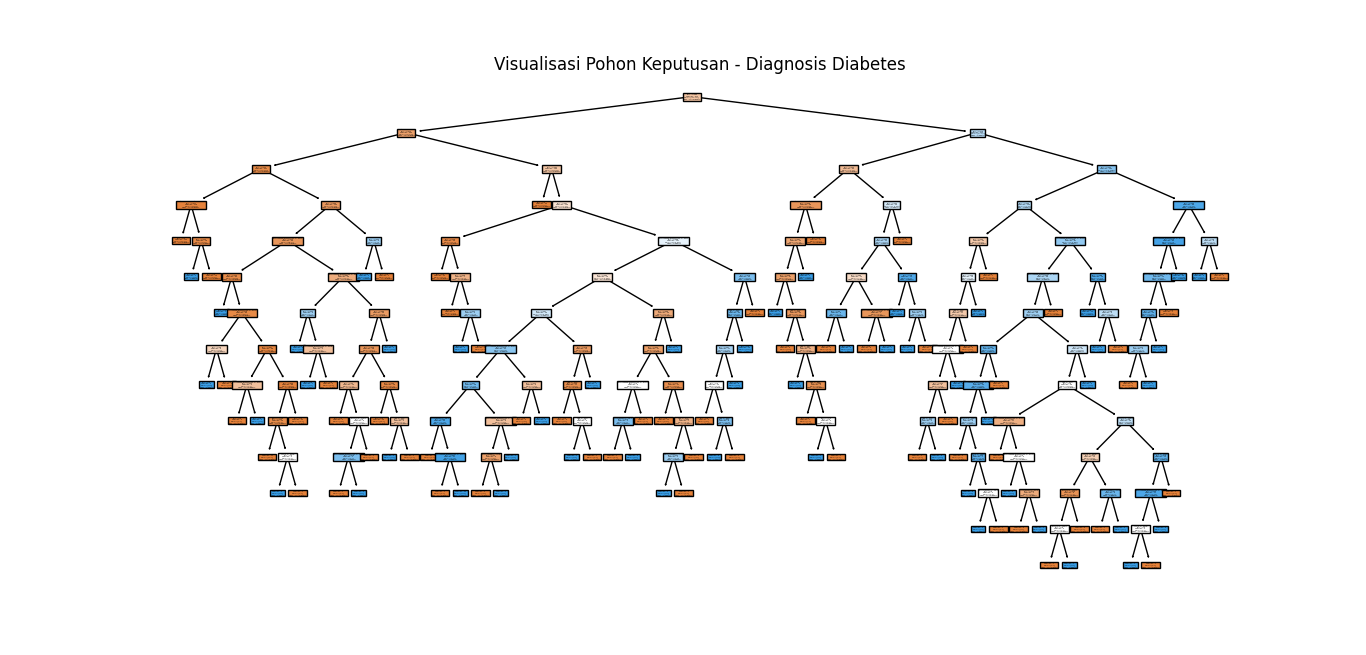
* **Precision Kelas 1 (Diabetes)**: 60%
* **Recall Kelas 1 (Diabetes)**: 69%

📌 **Interpretasi:**

* Model cukup baik dalam mendeteksi pasien diabetes (recall 69%), tapi masih ada yang salah diklasifikasikan.
* Model lebih akurat dalam mengenali yang *tidak diabetes* (kelas 0) dibandingkan yang *diabetes* (kelas 1).

📌 **Rekomendasi:**

* Gunakan **pruning** atau **depth limitation** untuk mencegah overfitting.
* Pertimbangkan teknik ensemble seperti **Random Forest** untuk hasil yang lebih stabil.
* Lakukan **cross-validation** untuk validasi yang lebih kuat.

**d. Visualisasi Pohon Keputusan**

### 📌 ****1. Struktur Umum Pohon****

* **Pohon dimulai dari akar (root)** di bagian paling atas.
* Tiap **node** menyatakan sebuah **kondisi atau aturan pemisahan data** berdasarkan fitur (contohnya Glucose <= 127.5).
* **Cabang (split)** menyatakan arah berdasarkan apakah kondisi itu benar (kiri) atau salah (kanan).
* **Daun (leaf node)** menunjukkan keputusan akhir, yaitu apakah data diklasifikasi sebagai **diabetes (class = 1) atau tidak (class = 0)**.

### 🎯 ****2. Isi Tiap Node (Kotak)****

Tiap node biasanya berisi:

* feature <= nilai: aturan pemisahan berdasarkan fitur tertentu.
* gini: ukuran impuritas (makin kecil, makin "murni" node tersebut).
* samples: jumlah data di node itu.
* value = [non\_diabetes, diabetes]: jumlah data per kelas.
* class: prediksi kelas dominan pada node.

### 🟦🟧 ****3. Warna Node****

* **Warna biru** = prediksi kelas **0 (Non-Diabetes)**.
* **Warna oranye** = prediksi kelas **1 (Diabetes)**.
* Semakin **gelap** warna, semakin **dominan** keputusan terhadap kelas itu.
* **Node paling atas** biasanya abu-abu muda karena mencakup seluruh data (masih campuran antara diabetes dan tidak).

### 🔍 ****4. Contoh Penafsiran Jalur****

Misalnya:

* Jika Glucose <= 127.5 → lanjut ke kiri.
* Jika BMI <= 30.1 → lanjut ke kiri lagi.
* Jika Age <= 25.5 → lanjut ke kiri lagi.
* ... hingga ke daun → class = 0 → artinya model memprediksi **tidak diabetes**.

Setiap jalur dari atas ke bawah menunjukkan **logika langkah demi langkah** dalam pengambilan keputusan.

### 📊 ****5. Kelebihan Visualisasi Ini****

* Memberikan **transparansi dan interpretabilitas**: kamu bisa tahu **mengapa** sebuah data diklasifikasikan sebagai diabetes.
* Bisa digunakan untuk menjelaskan keputusan ke pihak non-teknis, seperti dokter atau pengambil keputusan.