Nama: Ariel Modjo Satriawan

NIM: 231011403476

Kelas: 05 TPLE 004

Mata Kuliah: Machine Learning

1. Deskripsi Dataset

Dataset yang digunakan pada proyek ini adalah **Iris Dataset**, yaitu salah satu dataset klasik dalam bidang machine learning. Dataset ini berisi **150 data bunga iris**, dengan empat fitur utama dan satu variabel target.

Fitur-fitur dalam dataset adalah:

- **Sepal Length (cm)** — panjang kelopak bunga
- **Sepal Width (cm)** — lebar kelopak bunga
- **Petal Length (cm)** — panjang mahkota bunga
- **Petal Width (cm)** — lebar mahkota bunga
  Sedangkan **variabel target** adalah species, yang memiliki tiga kelas:

1. *Iris setosa*
2. *Iris versicolor*
3. *Iris virginica*

Tujuan analisis ini adalah **membangun model klasifikasi** untuk memprediksi jenis bunga berdasarkan keempat fitur di atas. Dataset ini dipilih karena sederhana, bersih, dan cocok digunakan untuk perbandingan performa algoritma klasifikasi yang berbeda.

---

2. Langkah-Langkah Analisis

a. Exploratory Data Analysis (EDA)

Tahapan awal dilakukan dengan melihat struktur dan distribusi data. Dari hasil eksplorasi:

- Tidak ditemukan nilai yang hilang (missing value).
- Setiap kelas memiliki jumlah data yang seimbang (masing-masing 50 data).
- Hubungan antar variabel menunjukkan pola yang cukup jelas. Misalnya, *petal length* dan *petal width* sangat berkorelasi dengan jenis bunga.

Visualisasi menggunakan *pairplot* dari seaborn menunjukkan bahwa *Iris setosa* mudah dipisahkan secara linear, sedangkan *versicolor* dan *virginica* sedikit tumpang tindih — ini akan menjadi tantangan kecil bagi model.

b. Preprocessing Data

Sebelum dilakukan pemodelan, dataset dibagi menjadi dua bagian:

- **80% data untuk pelatihan (training)**
- **20% data untuk pengujian (testing)**

Kemudian dilakukan **standarisasi** pada fitur menggunakan StandardScaler, agar semua fitur memiliki skala yang sama. Hal ini penting terutama untuk algoritma seperti Logistic Regression yang sensitif terhadap skala data.

Target variabel (species) diubah dari teks menjadi angka dengan LabelEncoder, menghasilkan label 0, 1, dan 2 untuk masing-masing kelas.

---

3. Model yang Digunakan

a. Logistic Regression

Model pertama adalah **Logistic Regression**, yaitu metode klasifikasi berbasis regresi linier yang memetakan probabilitas kelas menggunakan fungsi logistik (sigmoid).
Model ini bekerja dengan mencari batas linear terbaik yang memisahkan kelas-kelas data. Logistic Regression umumnya memberikan hasil baik pada data yang dapat dipisahkan secara linear.

Parameter utama:

LogisticRegression(max_iter=200)

Model dilatih menggunakan data yang telah distandarisasi.

b. Decision Tree

Model kedua adalah **Decision Tree Classifier**, yaitu metode berbasis struktur pohon yang membagi data ke dalam cabang-cabang berdasarkan nilai fitur. Setiap percabangan bertujuan meminimalkan impurity (ketidakmurnian kelas), menggunakan kriteria seperti *Gini Index* atau *Entropy*.

Model ini sangat interpretatif karena hasilnya bisa divisualisasikan dalam bentuk pohon keputusan. Namun, kelemahan utamanya adalah potensi *overfitting* pada dataset kecil.

Parameter utama:

DecisionTreeClassifier(random_state=42)

---

4. Evaluasi Model

Kedua model dievaluasi menggunakan beberapa metrik, yaitu:

- **Confusion Matrix**: menunjukkan jumlah prediksi benar dan salah untuk tiap kelas.
- **Accuracy**: proporsi prediksi benar dari seluruh data.
- **Precision, Recall, F1-score**: menggambarkan keseimbangan antara ketepatan dan kelengkapan model.
- **ROC Curve & AUC**: (jika relevan) mengukur kemampuan model membedakan antar kelas berdasarkan probabilitas.

Hasil evaluasi disajikan pada tabel berikut:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.967	0.968	0.967	0.967
Decision Tree	1.000	1.000	1.000	1.000

Dari confusion matrix diketahui bahwa:

- Logistic Regression hanya salah pada satu atau dua data dari 30 data uji.
- Decision Tree berhasil mengklasifikasikan seluruh data dengan benar.

---

5. Pembahasan

Secara umum, kedua model menunjukkan performa yang sangat baik pada dataset Iris. Hal ini dapat dijelaskan karena dataset ini memiliki pola yang jelas dan relatif mudah dipisahkan antar kelas.

Logistic Regression bekerja dengan sangat efisien dan memberikan hasil yang hampir sempurna. Model ini cocok untuk data dengan hubungan linear antar variabel.

Decision Tree, di sisi lain, mencapai akurasi sempurna karena mampu membentuk aturan yang sangat spesifik untuk memisahkan kelas berdasarkan fitur-fitur numerik. Namun, performa sempurna ini bisa menjadi tanda *overfitting* — artinya model terlalu menyesuaikan diri dengan data latih dan mungkin tidak sebaik itu pada data baru yang berbeda pola.

---

6. Kesimpulan

Berdasarkan hasil eksperimen, dapat disimpulkan bahwa:

1. **Kedua algoritma sama-sama cocok digunakan untuk dataset Iris**, dengan hasil evaluasi yang sangat tinggi.
2. **Decision Tree** memiliki akurasi lebih tinggi (100%) dibanding Logistic Regression (96,7%), namun perlu hati-hati terhadap potensi *overfitting*.
3. **Logistic Regression** lebih sederhana, cepat, dan stabil untuk digunakan pada dataset yang lebih besar dan kompleks.
4. Untuk kasus nyata, pemilihan model sebaiknya tidak hanya berdasarkan akurasi, tetapi juga memperhatikan interpretabilitas, kompleksitas model, serta ketersediaan data.

Dengan demikian, hasil ini menunjukkan pentingnya membandingkan beberapa algoritma untuk menemukan model terbaik sesuai karakteristik data yang digunakan.