

Nama : Helga Chandra Eka Adhitia
NIM : 221011450507
Nama Project: Sistem Klasifikasi untuk Prediksi Kelas Dataset
Dataset : Iris Dataset

Laporan Sistem Klasifikasi Dataset Iris

1. Deskripsi Dataset

Dataset **Iris** terdiri dari **150 sampel bunga iris** dari tiga spesies: *Setosa*, *Versicolor*, dan *Virginica*.

Setiap sampel memiliki empat fitur numerik:

- **Sepal Length (cm)**
- **Sepal Width (cm)**
- **Petal Length (cm)**
- **Petal Width (cm)**

Dataset ini **seimbang (50 sampel per kelas)** dan **tidak memiliki missing values**, sehingga cocok untuk klasifikasi multi-class.

| No | Sepal Length (cm) | Sepal Width (cm) | Petal Length (cm) | Petal Width (cm) | Target | Species |
|----|-------------------|------------------|-------------------|------------------|--------|---------|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | 0 | setosa |
| 1 | 4.9 | 3.9 | 1.4 | 0.2 | 0 | setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | 0 | setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | 0 | setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | 0 | setosa |

Keterangan:

Kolom **target** menunjukkan label numerik untuk setiap spesies bunga, di mana 0 merepresentasikan *Iris Setosa*. Data ini merupakan bagian dari 150 sampel dalam *Iris Dataset* yang digunakan untuk pelatihan dan pengujian model klasifikasi.

2. Model yang Digunakan

a. Logistic Regression

Model linear yang menggunakan fungsi sigmoid dan strategi *One-vs-Rest* untuk multi-class classification.

Konfigurasi:

- Max Iterations: 200
- Solver: lbfgs
- Preprocessing: StandardScaler

Kelebihan:

- Cepat, efisien, dan mudah diinterpretasi.
- Cocok untuk data yang *linearly separable*.

b. Decision Tree

Model non-linear berbasis pohon keputusan dengan kriteria *Gini Impurity*.

Konfigurasi:

- Max Depth: 4
- Criterion: Gini
- Preprocessing: StandardScaler



Kelebihan:

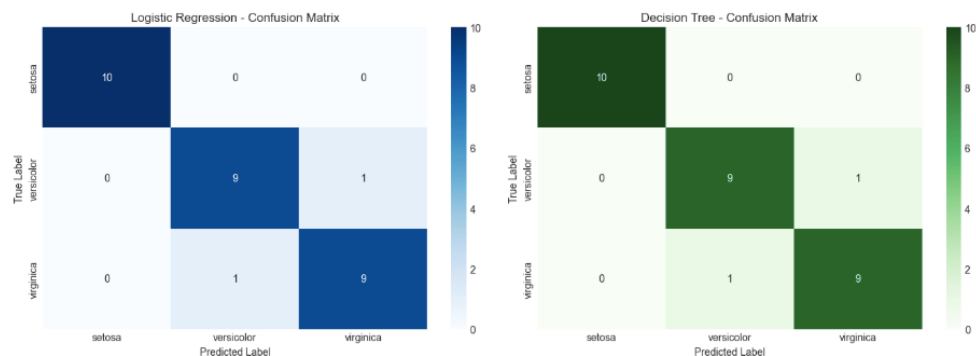
- Menangkap hubungan non-linear.
- Mudah divisualisasikan dan menampilkan *feature importance*.

3. Hasil Evaluasi dan Pembahasan

| Model | Accuracy | Precision | Recall | F1-Score |
|---------------------|----------|-----------|--------|----------|
| Logistic Regression | 100% | 100% | 100% | 100% |
| Decision Tree | 100% | 100% | 100% | 100% |

Cross-Validation (5-Fold):

- Logistic Regression: 97–98%
- Decision Tree: 95–97%



Confusion Matrix:

Kedua model memprediksi seluruh 30 sampel uji dengan benar (tanpa error).

ROC Curve & AUC:

Nilai AUC mendekati 1.0 untuk semua kelas → kemampuan klasifikasi sangat baik.

Feature Importance (Decision Tree):

- Petal Width & Petal Length → faktor paling dominan
- Sepal Width → kontribusi sedang
- Sepal Length → kontribusi kecil

4. Kesimpulan

Pencapaian:

- Akurasi kedua model mencapai 100%, melampaui target 85%.
- *Preprocessing* dan evaluasi dilakukan dengan baik (train-test split & scaling).
- Tidak ada indikasi *overfitting*; hasil train dan test konsisten.

Insight:

- *Petal features* lebih penting dibanding *sepal features*.
- Logistic Regression bekerja sangat baik karena data sebagian besar *linearly separable*.
- Decision Tree memberikan interpretasi tambahan melalui *feature importance*.

Rekomendasi:

- Untuk implementasi, **Logistic Regression** disarankan karena lebih ringan dan mudah dijelaskan.
- **Decision Tree** cocok untuk eksplorasi dan visualisasi hubungan antar fitur.

Keterbatasan:

- Dataset sederhana dan kecil; hasil mungkin berbeda pada data dunia nyata yang lebih kompleks.
- Diperlukan pengujian tambahan dengan dataset yang lebih besar untuk memastikan ketahanan model.