

UTS MECHINE LEARNING



OLEH :
TIJAN DARURI
(231011403108)
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA
UNIVERSITAS PAMULANG
TAHUN 2024/2025

1. Pendahuluan

Perkembangan teknologi *machine learning* telah memberikan dampak yang signifikan dalam bidang analisis data dan pengambilan keputusan. Salah satu penerapannya adalah pada masalah klasifikasi, yaitu proses memprediksi kelas atau kategori dari suatu data berdasarkan pola yang terdapat dalam data historis. Dalam konteks penelitian ini, digunakan **dataset Iris**, sebuah dataset klasik yang sering digunakan untuk pengenalan konsep klasifikasi dalam *data science* dan kecerdasan buatan.

Dataset Iris berisi informasi mengenai karakteristik morfologi tiga spesies bunga iris: *Iris setosa*, *Iris versicolor*, dan *Iris virginica*. Setiap data memiliki empat atribut numerik, yaitu **sepal length**, **sepal width**, **petal length**, dan **petal width**, yang kemudian digunakan untuk memprediksi jenis spesies bunga.

Tujuan utama penelitian ini adalah untuk:

1. Membangun model klasifikasi menggunakan dua algoritma *machine learning*, yaitu **Logistic Regression** dan **Decision Tree**.

2. Membandingkan kinerja kedua model berdasarkan metrik evaluasi seperti **accuracy**, **precision**, **recall**, dan **F1-score**.
3. Menarik kesimpulan mengenai model mana yang lebih efektif dalam mengklasifikasikan data bunga Iris.

2. Deskripsi Dataset

Dataset Iris pertama kali diperkenalkan oleh **Ronald A. Fisher** pada tahun 1936. Dataset ini terdiri dari:

- Jumlah sampel: **150 baris data**
- Jumlah fitur: **4 atribut numerik**
- Target (kelas): **3 spesies bunga Iris**

Fitur	Deskripsi
Sepal Length (cm)	Panjang kelopak bunga
Sepal Width (cm)	Lebar kelopak bunga
Petal Length (cm)	Panjang mahkota bunga
Petal Width (cm)	Lebar mahkota bunga

Distribusi kelas pada dataset adalah seimbang, masing-masing kelas (*setosa*, *versicolor*, *virginica*) memiliki 50 data. Data bersih dan tidak mengandung nilai hilang (*missing value*), sehingga tidak memerlukan proses imputasi.

3. Metodologi Penelitian

Langkah-langkah analisis dilakukan sebagai berikut:

Data dibagi menjadi **data latih (75%)** dan **data uji (25%)** menggunakan *stratified sampling* agar proporsi kelas tetap seimbang. Untuk algoritma yang sensitif terhadap skala data seperti Logistic Regression, dilakukan *feature scaling* menggunakan **StandardScaler**.

1. **Logistic Regression (One-vs-Rest)**
Model ini bekerja dengan membangun fungsi logistik untuk setiap kelas, kemudian memilih kelas dengan probabilitas tertinggi. Logistic Regression efektif untuk data yang memiliki batas pemisah linier.
2. **Decision Tree Classifier**
Model ini membangun pohon keputusan berdasarkan fitur yang memberikan Model dievaluasi menggunakan beberapa metrik umum:

- **Confusion Matrix** untuk melihat distribusi prediksi benar dan salah.
- **Accuracy, Precision, Recall, dan F1-score** sebagai ukuran kinerja keseluruhan.
- **ROC Curve (One-vs-Rest)** untuk menganalisis kemampuan model dalam membedakan kelas.

4. Hasil dan Pembahasan

Berdasarkan hasil pengujian pada data uji, diperoleh hasil sebagai berikut:

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)
Logistic Regression	0.816	0.828	0.818	0.821
Decision Tree	0.895	0.903	0.897	0.897

Analisis Hasil:

- Kedua model menunjukkan performa yang cukup baik, dengan akurasi di atas 80%.
- **Decision Tree** memiliki performa sedikit lebih baik dibanding Logistic Regression. Hal ini disebabkan karena Decision Tree mampu menangkap hubungan non-linier antar fitur yang tidak dapat ditangani secara optimal oleh Logistic Regression.
- Berdasarkan *confusion matrix*, kedua model dapat mengklasifikasikan *Iris setosa* dengan sangat baik, namun beberapa kesalahan prediksi terjadi antara *versicolor* dan *virginica* karena karakteristik morfologinya yang mirip.
- *ROC Curve* menunjukkan bahwa kedua model memiliki area di bawah kurva (AUC) yang tinggi, mengindikasikan kemampuan klasifikasi yang baik.

5. Kesimpulan

Dari hasil eksperimen dapat disimpulkan bahwa:

1. Kedua algoritma, **Logistic Regression** dan **Decision Tree**, mampu melakukan klasifikasi pada dataset Iris dengan performa yang baik.
2. **Decision Tree** memberikan hasil lebih tinggi dengan akurasi **89,47%**, dibandingkan Logistic Regression dengan **81,57%**.
3. Logistic Regression unggul dalam stabilitas dan interpretasi koefisien, sedangkan Decision Tree unggul dalam fleksibilitas dan menangani data non-linier.