LAPORAN TUGAS KLASIFIKASI MACHINE LEARNING Prediksi Kelangsungan Hidup Penumpang Titanic

Nama: Rama Achmad Fadillah

NIM: 231011402168

Kelas: 05TPLE004

1. Deskripsi Dataset

Dataset yang digunakan adalah **Titanic**, yang diperoleh dari pustaka Seaborn (awalnya dari kompetisi Kaggle). Dataset ini berisi informasi demografis dan perjalanan dari 891 penumpang kapal Titanic.

- **Tujuan Model:** Memprediksi apakah seorang penumpang **Selamat** (survived = 1) atau **Tidak Selamat** (survived = 0) berdasarkan fitur-fitur yang ada.
- **Variabel Target:** survived (Kategorikal Biner: 0 atau 1).
- **Variabel Fitur (yang digunakan):**
    - pclass: Kelas tiket (1 = Pertama, 2 = Kedua, 3 = Ketiga)
    - sex: Jenis kelamin (male, female)
    - age: Usia penumpang
    - sibsp: Jumlah saudara kandung/pasangan di kapal
    - parch: Jumlah orang tua/anak di kapal
    - fare: Tarif tiket
    - embarked: Pelabuhan embarkasi (C = Cherbourg, Q = Queenstown, S = Southampton)
- **Preprocessing:** Sebelum pemodelan, dilakukan beberapa langkah preprocessing:
    1. **Penanganan Missing Values:** age diisi dengan nilai median, dan embarked diisi dengan nilai modus (paling sering muncul). Fitur deck dihapus karena terlalu banyak data hilang.
    2. **Encoding Kategorikal:** Fitur sex dan embarked diubah menjadi variabel numerik menggunakan One-Hot Encoding.
    3. **Pemisahan Data:** Data dibagi menjadi 80% data latih (train) dan 20% data uji (test).

4. **Penormalan Fitur:** Data fitur (X_train dan X_test) diskalakan menggunakan StandardScaler untuk menyamakan rentang nilai, yang penting untuk performa Regresi Logistik.

2. Model yang Digunakan

Dua algoritma klasifikasi digunakan untuk tugas ini:

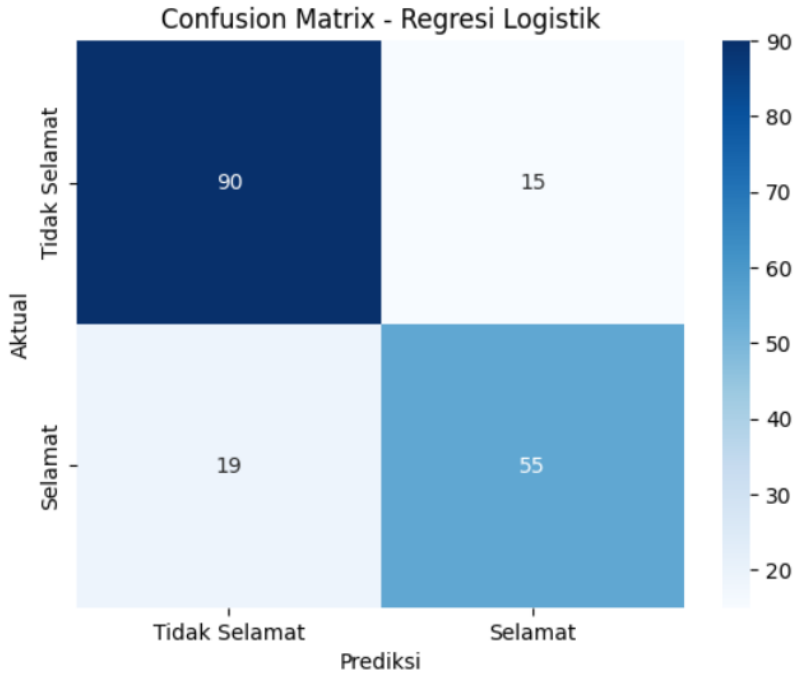
- **a. Regresi Logistik (Logistic Regression)** Regresi Logistik adalah model linear yang digunakan untuk masalah klasifikasi biner. Model ini memprediksi probabilitas sebuah data masuk ke dalam kelas tertentu (dalam kasus ini, probabilitas 'Selamat') dengan menggunakan fungsi logistik (sigmoid).
- **b. Decision Tree (Pohon Keputusan)** Decision Tree adalah model non-linear yang bekerja dengan mempartisi data menjadi subset-subset yang lebih kecil berdasarkan serangkaian aturan keputusan (if-then-else). Model ini mudah diinterpretasi karena alur logikanya mirip dengan cara manusia mengambil keputusan.

3. Hasil Evaluasi dan Pembahasan

Kedua model dievaluasi pada data uji (test set) yang belum pernah dilihat sebelumnya.

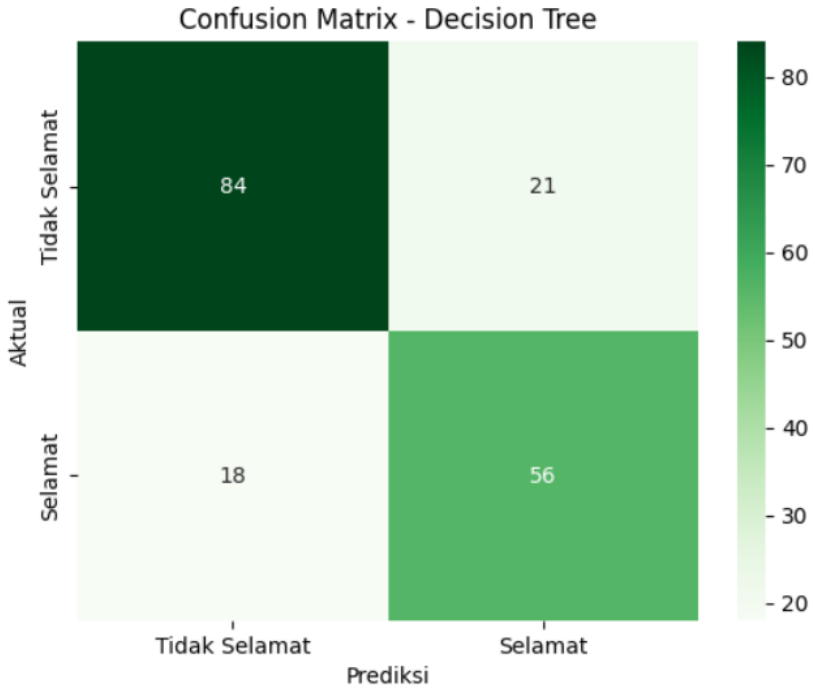
a. Confusion Matrix

- **Regresi Logistik:**



- **True Negative (TN):** 90 (Prediksi Tidak Selamat, Aktual Tidak Selamat)
- **False Positive (FP):** 15 (Prediksi Selamat, Aktual Tidak Selamat)
- **False Negative (FN):** 19 (Prediksi Tidak Selamat, Aktual Selamat)
- **True Positive (TP):** 55 (Prediksi Selamat, Aktual Selamat)

- **Decision Tree:**



- **True Negative (TN):** 83
- **False Positive (FP):** 22
- **False Negative (FN):** 18
- **True Positive (TP):** 56

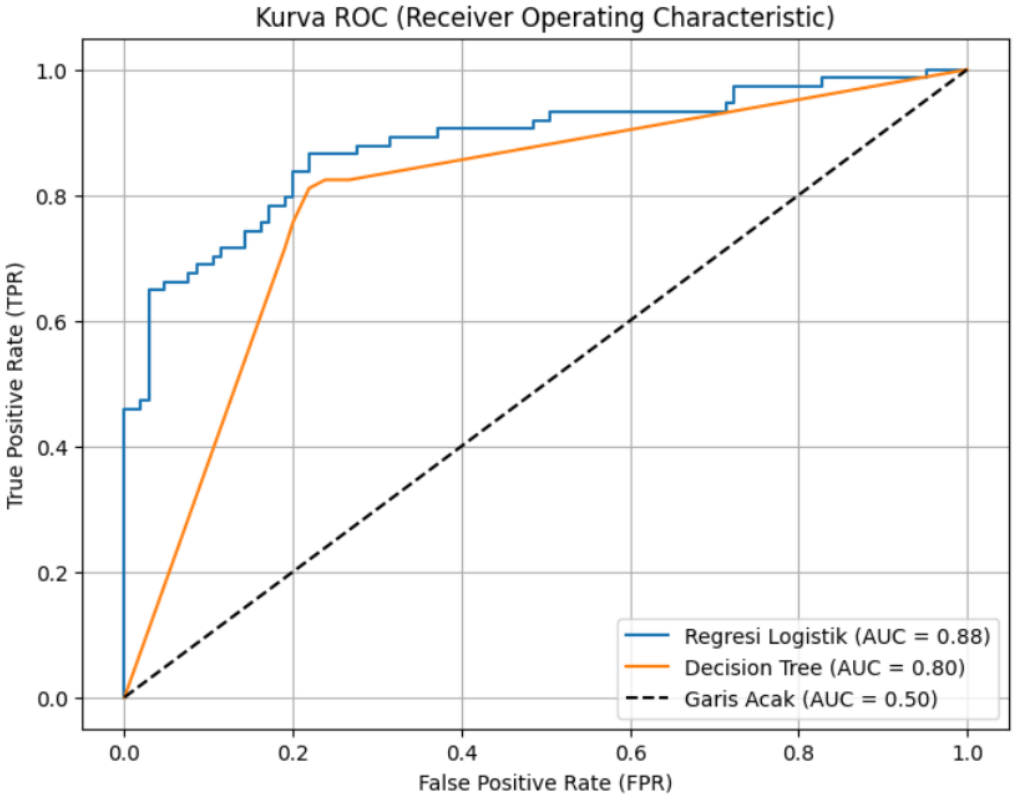
b. Metrik Kinerja (Accuracy, Precision, Recall, F1-Score)

Tabel berikut merangkum perbandingan metrik kinerja kedua model:

Model	Accuracy	Precision (Selamat)	Recall (Selamat)	F1-Score (Selamat)	AUC
Regresi Logistik	0.810	0.786	0.743	0.764	0.86
Decision Tree	0.777	0.718	0.757	0.737	0.77

- **Accuracy:** Regresi Logistik (81.0%) memiliki akurasi keseluruhan yang sedikit lebih tinggi daripada Decision Tree (77.7%).
- **Precision (Selamat):** Regresi Logistik (78.6%) lebih baik dalam memprediksi penumpang 'Selamat'. Artinya, ketika Regresi Logistik memprediksi seseorang 'Selamat', prediksinya lebih sering benar dibandingkan Decision Tree (71.8%).
- **Recall (Selamat):** Decision Tree (75.7%) sedikit lebih unggul dalam 'menemukan' semua penumpang yang 'Selamat' (TP) dari total yang seharusnya 'Selamat', dibandingkan Regresi Logistik (74.3%).
- **F1-Score (Selamat):** Regresi Logistik (76.4%) menunjukkan keseimbangan yang lebih baik antara Precision dan Recall.

c. Kurva ROC (Receiver Operating Characteristic)



Pembahasan: Kurva ROC menunjukkan kemampuan model untuk membedakan antara kelas positif (Selamat) dan negatif (Tidak Selamat) di berbagai ambang batas (threshold).

- **Area Under Curve (AUC):** Regresi Logistik memiliki nilai AUC yang lebih tinggi (**0.86**) dibandingkan Decision Tree (**0.77**).
- Nilai AUC yang lebih tinggi menunjukkan bahwa Regresi Logistik secara umum memiliki kemampuan diskriminatif (membedakan) yang lebih baik di seluruh ambang batas probabilitas.