UTS MACHINE LEARNING

Nama : Fernanda Harjoki Putra

Nim : 231011400229

Kelas : 05TPLE005

DESKRIPSI DATASET

Dataset yang digunakan dalam proyek ini adalah Titanic.csv, yang berisi informasi tentang penumpang kapal Titanic dan status keselamatan mereka (selamat atau tidak selamat). Dataset ini memiliki 891 baris dan 12 kolom, dengan variabel target bernama Survived yang bersifat kategorikal — bernilai 1 jika penumpang selamat dan 0 jika tidak selamat.

Struktur Dataset :

- **PassengerId** – ID unik penumpang
- **Pclass** – kelas tiket (1, 2, atau 3)
- **Name** – nama penumpang
- **Sex** – jenis kelamin
- **Age** – usia penumpang
- **SibSp** – jumlah saudara atau pasangan di kapal
- **Parch** – jumlah orang tua atau anak di kapal
- **Ticket** – nomor tiket
- **Fare** – harga tiket
- **Cabin** – nomor kabin
- **Embarked** – pelabuhan tempat naik (C = Cherbourg, Q = Queenstown, S = Southampton)

Hasil EDA (Exploratory Data Analysis):

1. **Jumlah data dan tipe kolom:**
   Terdapat 12 fitur dengan tipe data numerik dan kategorikal.

2. **Data yang hilang (missing values):**
   Kolom Age: 177 nilai kosong
   Kolom Cabin: 687 nilai kosong
   Kolom Embarked: 2 nilai kosong

3. **Langkah preprocessing:**
   Age diisi menggunakan nilai **rata-rata (mean)**
   Embarked diisi dengan **modus (S)**
   Cabin dihapus karena banyaknya data kosong

4. **Distribusi target (Survived):**
   Tidak selamat: 61.6%
   Selamat: 38.4%

Insight Awal:

- **Perempuan** memiliki tingkat keselamatan lebih tinggi (74.2%) dibanding **laki-laki** (18.9%).
- Berdasarkan kelas, **kelas 1** memiliki peluang selamat tertinggi (62.9%), diikuti kelas 2 (47.3%) dan kelas 3 (24.2%).

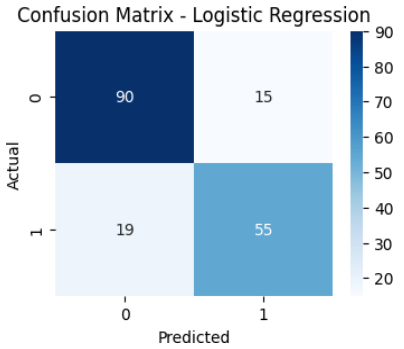
MODEL YANG DIGUNAKAN

Penelitian ini menggunakan dua algoritma klasifikasi untuk memprediksi keselamatan penumpang Titanic, yaitu **Logistic Regression** dan **Decision Tree**.

Logistic Regression

Logistic Regression merupakan model berbasis regresi linier yang digunakan untuk memprediksi probabilitas suatu kejadian (dalam hal ini, penumpang selamat atau tidak).

Model ini mengubah output linier menjadi probabilitas melalui fungsi **sigmoid (logistik)**.



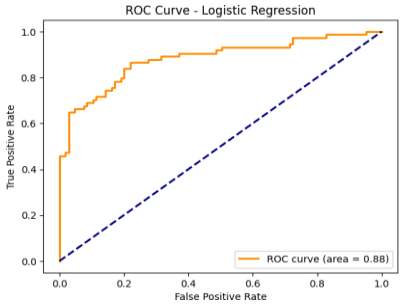
Keunggulan:

- Mampu menangani variabel kategorikal dan numerik
- Stabil terhadap noise
- Mudah ditafsirkan dan memiliki generalisasi yang baik

Hasil sementara:

- Akurasi: **81%**
- ROC AUC: **0.88**
- Model mampu mengenali pola keselamatan dengan baik dan tidak mengalami overfitting.

Decision Tree



Decision Tree adalah algoritma yang membentuk struktur pohon keputusan berdasarkan nilai fitur yang paling berpengaruh terhadap target.
Cabang pohon menunjukkan keputusan yang diambil untuk memisahkan data dalam kelompok yang lebih homogen.

Keunggulan:

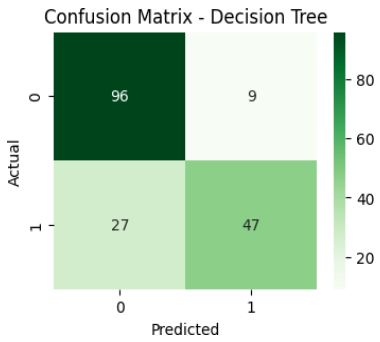
- Sederhana dan mudah divisualisasikan
- Menangkap hubungan non-linear antar fitur
- Tidak memerlukan normalisasi data

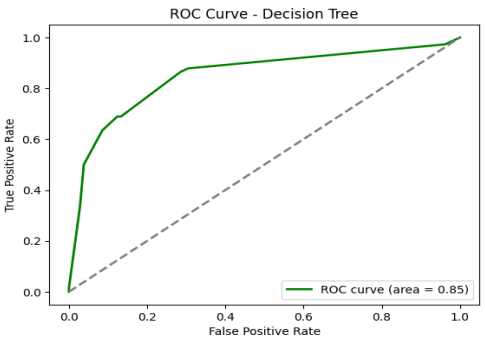
Kelemahan:

- Mudah mengalami **overfitting**
- Performa bisa menurun pada data dengan banyak noise

Hasil sementara:

- Akurasi: **79.9%**
- ROC AUC: **0.85**
- Decision Tree menghasilkan interpretasi visual yang mudah dipahami namun sedikit lebih rendah kinerjanya dibanding Logistic Regression.





HASIL EVALUASI DAN PEMBAHASAN

Evaluasi model dilakukan menggunakan data uji sebanyak 179 baris dengan metrik **Accuracy**, **Precision**, **Recall**, **F1-Score**, dan **ROC AUC** untuk menilai performa model **Logistic Regression** dan **Decision Tree**.

1. Logistic Regression
Model ini menunjukkan performa yang baik dengan hasil: akurasi **81%**, precision **0.79**, recall **0.74**, F1-score **0.76**, dan ROC AUC **0.88**. Nilai AUC yang tinggi menandakan kemampuan model dalam membedakan kelas "selamat" dan "tidak selamat" secara sangat baik. Logistic Regression memiliki keseimbangan antara precision dan recall, serta generalisasi yang stabil tanpa tanda overfitting. Berdasarkan confusion matrix, sebagian besar penumpang berhasil diprediksi dengan benar.

2. Decision Tree
Model Decision Tree menghasilkan akurasi **79.9%**, precision **0.84**, recall **0.64**, F1-score **0.72**, dan ROC AUC **0.85**. Model ini memiliki precision yang lebih tinggi pada kelas "selamat", namun recall lebih rendah sehingga beberapa penumpang selamat tidak terdeteksi. Kelebihan utama Decision Tree adalah kemampuannya memberikan interpretasi yang jelas dalam bentuk diagram pohon keputusan, misalnya: *penumpang perempuan di kelas 1 memiliki peluang selamat lebih tinggi.*

3. Perbandingan Model

Model	Akurasi	ROC AUC	Ciri Utama
Logistic Regression	**0.81**	**0.88**	Linear, stabil, generalisasi baik
Decision Tree	**0.80**	**0.85**	Non-linear, mudah dijelaskan, rawan overfitting

Perbedaan performa kedua model relatif kecil (±1%), namun Logistic Regression unggul pada hampir semua metrik evaluasi. Decision Tree tetap berguna untuk pemahaman visual dan eksplorasi pola, tetapi cenderung lebih sensitif terhadap data outlier.

KESIMPULAN
Kedua model sama-sama mampu memprediksi keselamatan penumpang Titanic dengan akurasi tinggi. Namun, **Logistic Regression** merupakan model terbaik karena memiliki **akurasi tertinggi (81%)**, **ROC AUC tertinggi (0.88)**, serta performa seimbang antara precision dan recall. Sementara **Decision Tree** cocok untuk analisis interpretatif berkat visualisasi pohonnya. Secara keseluruhan, Logistic Regression direkomendasikan karena lebih stabil dan memiliki kemampuan generalisasi yang baik.