

**Analisis Klasifikasi Dataset Kanker Payudara
Menggunakan Algoritma Machine Learning**



Oleh:

Thasya Mulia (1103201208)

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

A. Pendahuluan

Tujuan dari laporan ini adalah untuk melakukan analisis klasifikasi pada dataset kanker payudara menggunakan beberapa algoritma pembelajaran mesin yaitu Decision Tree, Random Forest, dan Self-Training. Dataset yang digunakan dalam analisis ini adalah Breast Cancer Wisconsin (Diagnostic) Dataset yang disediakan oleh scikit-learn.

B. Data Preparation

Pertama-tama, kita memuat dataset menggunakan fungsi `load_breast_cancer()` dari library scikit-learn. Kemudian, dataset tersebut dikonversi menjadi pandas DataFrame dengan menambahkan kolom target.

Selanjutnya, kita melakukan eksplorasi awal pada data dengan menggunakan pairplot dari seaborn. Pairplot ini memungkinkan kita untuk memvisualisasikan hubungan antara beberapa fitur dalam dataset dengan membaginya berdasarkan kelas target (malignant atau benign). Dari visualisasi ini, kita dapat melihat adanya beberapa fitur yang memiliki perbedaan distribusi antara kelas malignant dan benign.

C. Pelatihan Model

Selanjutnya, kita membagi dataset menjadi data pelatihan dan data pengujian menggunakan fungsi `train_test_split()` dari scikit-learn. Setelah itu, kita melatih model Decision Tree, Random Forest, dan Self-Training pada data pelatihan.

Pada Decision Tree dan Random Forest, kita menggunakan metrik akurasi untuk mengevaluasi kinerja model pada data pengujian. Kedua model tersebut mampu menghasilkan akurasi yang tinggi, yaitu sekitar 0.96 untuk Decision Tree dan 0.97 untuk Random Forest.

Pada Self-Training, kita menggunakan `SelfTrainingClassifier` dari scikit-learn dengan Decision Tree sebagai base classifier. Dalam hal ini, kita menggunakan metrik akurasi untuk mengevaluasi kinerja model. Hasilnya menunjukkan bahwa Self-Training mampu mencapai akurasi sekitar 0.94, meskipun hasil ini sedikit lebih rendah dibandingkan dengan hasil yang diperoleh oleh Decision Tree dan Random Forest.

D. Kesimpulan

Dari analisis klasifikasi yang dilakukan pada dataset kanker payudara menggunakan Decision Tree, Random Forest, dan Self-Training, dapat disimpulkan bahwa algoritma pembelajaran mesin mampu memprediksi kelas target dengan akurasi yang cukup tinggi. Namun, dalam memilih model yang paling cocok untuk dataset ini, perlu diperhatikan faktor-faktor seperti biaya dan risiko kesalahan yang dapat mempengaruhi keputusan final.