Technical Report: Klasifikasi Kanker Payudara dengan Model Decision Tree, Random Forest dan Self-Training.

Pada *Technical Report* ini akan dibahas langkah-langkah mengklasifikasikan kanker payudara menggunakan beberapa model diantaranya adalah *Decision Tree, Random Forest dan Self-Training*. Dataset yang digunakan adalah sklearn.datasets.load_breast_cancer yang berisi 569 sampel tumor ganas dan jinak dengan 30 parameter untuk setiap sampelnya.

Tahap Pertama: Import Library dan Load Dataset

Pada bagian ini, dilakukan import beberapa library yang dibutuhkan untuk pengolahan data seperti numpy, pandas, matplotlib, seaborn, dan beberapa library dari sklearn. Setelah itu, dilakukan load dataset menggunakan library sklearn.datasets.load_breast_cancer.

Tahap Kedua: Eksplorasi Data

Pada bagian ini, dilakukan beberapa eksplorasi data seperti melihat informasi umum dari data seperti jumlah baris dan kolom, tipe data dari masing-masing kolom, serta statistik deskriptif dari masing-masing kolom. Selanjutnya, dilakukan visualisasi data menggunakan library seaborn untuk melihat distribusi data dari masing-masing kolom serta hubungan antar kolom.

Tahap Ketiga: Pemisahan Data

Pada bagian ini, dilakukan pemisahan data menjadi data train dan data test menggunakan train_test_split dengan proporsi data train 80% dan data test sebesar 20%. Selanjutnya, dilakukan pemisahan data train menjadi labeled data dan unlabeled data dengan menyimpan beberapa data label sebagai data unlabeled.

Tahap Keempat: Modeling dan Evaluasi

Pada bagian ini, dilakukan pemodelan dan evaluasi menggunakan Decision Tree, Random Forest, dan Self-Training. Pertama, dilakukan pemodelan Decision Tree menggunakan DecisionTreeClassifier dan evaluasi menggunakan accuracy_score. Selanjutnya, dilakukan pemodelan Random Forest menggunakan RandomForestClassifier dan evaluasi menggunakan accuracy_score. Terakhir, dilakukan pemodelan Self-Training menggunakan SelfTrainingClassifier dengan DecisionTreeClassifier sebagai model classifier dan evaluasi menggunakan accuracy_score.

Tahap Kelima: Visualisasi Model

Pada bagian ini, dilakukan visualisasi model Decision Tree dan Random Forest menggunakan library graphviz. Selain itu, dilakukan visualisasi fitur terpenting dari model Random Forest menggunakan library seaborn.

Kesimpulan

Berdasarkan hasil evaluasi, ditemukan bahwa model Random Forest memiliki akurasi yang lebih tinggi dibandingkan dengan Decision Tree dan Self-Training. Selain itu, fitur terpenting dari model Random Forest adalah mean concave points dan worst perimeter. Visualisasi model Decision Tree dan Random Forest dapat membantu dalam pemahaman mengenai bagaimana model bekerja dan fitur-fitur apa yang digunakan dalam pengambilan keputusan.