Technical Report Breast Cancer Dataset Menggunakan Decision Tree, Random Forest, dan Self-Training

Oleh :

Muhamad Miftah Rizaldi Ruswandi (1103204196)

Mata Kuliah Machine Learning

Tahun 2023

Tujuan dari laporan Teknik ini adalah untuk mempresentasikan implementasi model klasifikasi pada Breast Cancer Dataset menggunakan Decision Tree, Random Forest, dan Self-Training Dataset yang digunakan adalah dataset bawaan scikit-learn, yaitu dataset Breast Cancer Wisconsin.

Data Prepocessing

Pertama-tama, dataset dimuat menggunakan '**load_breast_cancer'** dari sckit-learn. Selanjutnya, dataset dikonversi menjadi pandat DataFrame dan diatur kolomnya dengan '**pd.DataFrame'**. Kemudian, kolom target ditambahkan ke dataset menggunakan '**np.c'**. Data dipisahkan menjadi dua bagian, yaitu fitur dan tagert.

Eksplorasi Data

Untuk memvisualisasikan dataset, diguanakn Seaborn untuk membuat plot padangan (pair plot) menggunakan '**sns.pairplot()'**. Pada plot pasangan, nilai-nilai fitur dari pasien dikombinasikan dalam plot. Kolom target digunakan sebagai variable hue, sehingga plot pasangan diwarnai sesuai dengan kelar target.

Model Klasifikasi

Data selanjutnya dibagi menjadi set latih dan set uji dengan **'train_test_split'** dari scikit-learn. Kemudian, model Decision Tree dan Random Forest diaplikasikan pada set latih menggunakan '**DescisionTreeClassifier'** dan '**RandomForestClassifier'**. Untuk masing-masing model, nilai akurasi dihitung pada set uji menggunakan '**accuracu_score'** dari scikit-learn.

Self-Training

Selanjutnya, model Self-Training diaplikasikan pada set latih dengan menggunakan **'SelfTrainingClassifier'** dari scikit-learn. Model ini menggunakan Decision Tree sebagai base sclassifier dan maksimal iterasi yang diatur adalah 50. Seteleah dilatih, model ini diuji pada set uji dan akurasinya dihitung dengan menggunakan '**accuracy_score'**.

Kesimpulan

Dari hasil eksperimen ini, ditemukan bahwa model Random Forest memiliki akurasi paling tinggi pada dataset Breast Cancer Wisconsin. Model Decision Tree memiliki akurasi yang sedikit lebih rendah dibandingkan dengan Random Forest. Model Self-Training memiliki akurasi yang sama dengan model Decision Tree. Dalam hal ini, pengguna Teknik Self-Training tidak menghasilkan peningkatan signifikan pada akurasi model Decision Tree.

Oleh karena itu, dapat disimpulkan bahwa model Random Forest adalah model yang paling cocok untuk melakukan klasifikasi pada dataset Breat Cancer Wisconsin. Namun, model Decision Tree dan Self-Trainning juga dapat menjadi alternatif pilihan jika dibutuhkan model yang lebih sederhana dan interpretative.