Technical Report

Analisis Data Kanker Payudara Menggunakan Decision Tree, Random Forest, dan Self-Training Method

Pendahuluan

Kanker payudara adalah jenis kanker yang paling umum terjadi pada wanita. Seiring dengan perkembangan teknologi, metode pengobatan dan prediksi kanker payudara semakin ditingkatkan. Oleh karena itu, dalam laporan ini, akan dibahas penggunaan berbagai metode machine learning untuk menganalisis data kanker payudara, terutama menggunakan decision tree, random forest, dan self-training method.

Metode

Pertama-tama, dataset kanker payudara dimuat menggunakan fungsi load_breast_cancer dari paket Scikit-learn dan dikonversi menjadi pandas dataframe. Setelah itu, dilakukan exploratory data analysis dengan membuat scatter plot dan heatmap dari fitur-fitur dalam dataset. Berikutnya, dataset dibagi menjadi data training dan data testing menggunakan fungsi train_test_split. Kemudian dilakukan pembangunan model decision tree classifier dan dilakukan pruning menggunakan cost complexity pruning untuk meningkatkan performa pada data testing. Selain itu, juga dilakukan pembangunan model random forest classifier dan perhitungan feature importance menggunakan permutation importance. Terakhir, dilakukan pembangunan model self-training classifier dengan menggunakan base classifier support vector machine dan dilakukan pelatihan secara iteratif pada data yang terlabel dan tidak terlabel untuk meningkatkan akurasi pada data testing.

Hasil Pada tahap exploratory data analysis, ditemukan bahwa beberapa fitur pada dataset memiliki korelasi yang cukup kuat, seperti mean radius dan mean perimeter. Selain itu, scatter plot menunjukkan adanya perbedaan yang signifikan antara tumor jinak dan ganas. Berdasarkan hasil dari pembangunan model decision tree, terlihat bahwa pruning dapat meningkatkan performa model pada data testing. Model random forest menunjukkan bahwa feature importance tertinggi adalah mean concave points. Hasil dari pembangunan model self-training classifier menunjukkan peningkatan akurasi pada data testing dengan menggunakan pelatihan iteratif pada data terlabel dan tidak terlabel.

Kesimpulan

Dalam laporan ini, telah dilakukan analisis data kanker payudara menggunakan decision tree, random forest, dan self-training method. Dari hasil yang diperoleh, dapat disimpulkan bahwa penggunaan metode machine learning dapat membantu dalam prediksi kanker payudara. Selain itu, self-training method juga terbukti dapat meningkatkan akurasi pada data testing dengan menggunakan pelatihan iteratif pada data terlabel dan tidak terlabel. Oleh karena itu, diharapkan bahwa hasil dari laporan ini dapat memberikan kontribusi dalam pengembangan metode prediksi kanker payudara di masa yang akan datang.