TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

Diajukan untuk memenuhi tugas pengganti Ujian Tengah Semester (UTS)
pada mata kuliah Machine Learning

Disusun oleh :

Muhammad Rizky Pradhitia - 1103204192

PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS TEKNIK ELEKTRO
UNIVERSITAS TELKOM
2023

I. Pendahuluan

Kanker payudara adalah tumor ganas yang berasal dari sel-sel payudara. Deteksi dini dan diagnosis yang akurat dari kanker payudara sangat penting untuk pengobatan yang efektif dan hasil yang lebih baik bagi pasien. Algoritma pembelajaran mesin telah banyak digunakan dalam penelitian medis untuk mengklasifikasikan kanker payudara berdasarkan berbagai fitur yang diekstraksi dari data medis.

Dalam laporan teknis ini, disajikan pendekatan pembelajaran mesin untuk klasifikasi kanker payudara menggunakan dataset kanker payudara dari perpustakaan sklearn. Dataset ini berisi berbagai fitur yang diekstraksi dari pasien kanker payudara, seperti rata-rata radius, rata-rata tekstur, rata-rata keliling, rata-rata luas, dan rata-rata kehalusan, beserta label target yang menunjukkan apakah tumor tersebut ganas (1) atau jinak (0).

II. Persiapan Data

Pertama, memuat dataset kanker payudara menggunakan fungsi load_breast_cancer() dari modul sklearn.datasets. Argumen as_frame=True memungkinkan untuk memuat data ke dalam DataFrame pandas untuk memudahkan manipulasi data. kemudian melakukan eksplorasi awal data dan visualisasi untuk mendapatkan wawasan tentang dataset.

menggambarkan distribusi variabel target menggunakan countplot, yang menunjukkan bahwa dataset seimbang dengan proporsi kasus ganas dan jinak yang cukup seimbang. membuat pairplot untuk memvisualisasikan hubungan antara fitur-fitur yang berbeda dan variabel target, dan matriks korelasi heatmap untuk memahami korelasi antara fitur-fitur.

III. Pra-Pemrosesan Data

Selanjutnya, membagi data menjadi set pelatihan dan pengujian menggunakan fungsi train_test_split() dari modul sklearn.model_selection, dengan 80% data digunakan untuk pelatihan dan 20% untuk pengujian, menghapus variabel target dari matriks fitur dalam set pelatihan dan pengujian, dan menyimpannya terpisah sebagai vektor target.

IV. Pembangunan Model

dimemulai dengan membuat klasifikasi pohon keputusan menggunakan kelas DecisionTreeClassifier() dari modul sklearn.tree. menyesuaikan klasifikasi tersebut dengan data pelatihan menggunakan metode fit(), dan kemudian mencari nilai optimal untuk parameter kompleksitas biaya (ccp_alpha) menggunakan pemangkasan kompleksitas biaya. Dan mengulanginya untuk berbagai nilai ccp_alpha dan melatih klasifikasi pohon keputusan dengan tingkat pemangkasan yang bervariasi, menyimpan akurasi pelatihan dan pengujian mereka untuk setiap nilai ccp_alpha.

Selanjutnya, membuat klasifikasi pohon keputusan yang telah dipangkas dengan nilai ccp_alpha optimal yang ditemukan sebelumnya, dan mengevaluasi performanya pada data pengujian. Disini juga menghitung akurasi pohon keputusan yang telah dipangkas menggunakan fungsi accuracy_score() dari modul sklearn.metrics, dan memvisualisasikan pohon keputusan yang telah dipangkas menggunakan fungsi plot_tree() dari modul sklearn.tree.Kemudian membangun random forest menggunakan kelas RandomForestClassifier() dari modul sklearn.ensemble. menyesuaikan klasifikasi tersebut dengan data pelatihan menggunakan metode fit(), dan kemudian melakukan prediksi pada data pengujian menggunakan metode predict(). Untuk menghitung akurasi hutan random menggunakan fungsi accuracy_score() dari modul sklearn.metrics untuk mengevaluasi kinerjanya.

Selanjutnya, membangun klasifikasi Support Vector Machine (SVM) menggunakan kelas SVC() dari modul sklearn.svm. Kami menyesuaikan klasifikasi SVM dengan data pelatihan menggunakan metode fit(), dan melakukan prediksi pada data pengujian menggunakan metode predict().menghitung akurasi SVM menggunakan fungsi accuracy_score() untuk mengevaluasi kinerjanya.

Selain itu, kami juga membangun klasifikasi K-Nearest Neighbors (KNN) menggunakan kelas KNeighborsClassifier() dari modul sklearn.neighbors. menyesuaikan klasifikasi KNN dengan data pelatihan menggunakan metode fit(), dan melakukan prediksi pada data pengujian menggunakan metode predict(). menghitung akurasi KNN menggunakan fungsi accuracy_score() untuk mengevaluasi kinerjanya.

V. Hasil dan Evaluasi Algoritma

Di algoritma in membandingkan akurasi ketiga model yang telah dibangun, yaitu pohon keputusan yang telah dipangkas, hutan random, SVM, dan KNN. disini juga melakukan evaluasi performa model menggunakan metrik evaluasi lainnya seperti presisi, recall, dan F1-score untuk mendapatkan pemahaman yang lebih lengkap tentang kinerja model.

Hasil eksperimen menunjukkan bahwa hutan random memiliki akurasi yang paling tinggi, diikuti oleh SVM, KNN, dan pohon keputusan yang telah dipangkas. Namun, performa model dapat bervariasi tergantung pada dataset dan pemilihan parameter yang optimal. Oleh karena itu, tuning parameter dan validasi silang dapat dilakukan untuk meningkatkanperformamodel.

VI. Kesimpulan

Dalam laporan teknis ini, Kode melakukan analisis data eksploratori, implementasi algoritma, dan evaluasi algoritma. Algoritma yang diimplementasikan termasuk decision tree classifier, pruned decision tree classifier, random forest classifier, dan self-training classifier. dan telah membangun dan membandingkan beberapa model pembelajaran mesin untuk klasifikasi kanker payudara berdasarkan fitur-fitur dari data medis. Hasil eksperimen menunjukkan bahwa hutan random, SVM, dan KNN memiliki akurasi yang lebih tinggi dibandingkan dengan pohon keputusan yang telah dipangkas. Namun, pemilihan model terbaik masih bergantung pada karakteristik dataset dan kebutuhan aplikasi yang spesifik. Penggunaan algoritma pembelajaran mesin dalam diagnosis kanker payudara dapat membantu dalam deteksi dini dan pengobatan yang efektif, namun perlu validasi