TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

Mohammad Rizki Ramdhan 1103204126

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

1. Introduction

Kanker payudara adalah jenis kanker yang berkembang dari jaringan payudara. Deteksi dini dan klasifikasi kanker payudara sangat penting untuk pengobatan yang efektif. Algoritme pembelajaran mesin dapat digunakan untuk mengklasifikasikan kanker payudara dan memprediksi hasilnya. Dalam laporan ini, kami akan menggunakan dataset Kanker Payudara untuk melatih dan mengevaluasi tiga pengklasifikasi yang berbeda, yaitu Pengklasifikasi Pohon Keputusan, Pengklasifikasi Hutan Acak, dan Pengklasifikasi Pelatihan Mandiri.

2. Data Preprocessing

Pertama, dilakukan impor library yang dibutuhkan, seperti numpy, pandas, seaborn, matplotlib, dan scikit-learn. Kemudian, dataset Breast Cancer di-load menggunakan scikit-learn, dan dikonversi menjadi DataFrame menggunakan library pandas. Data kemudian dijelajahi menggunakan Seaborn dengan membuat pairplot untuk menampilkan korelasi antar fitur dalam dataset.

Setelah itu, dataset dibagi menjadi data latih dan data uji menggunakan train_test_split dari scikit-learn. Tiga model klasifikasi kemudian dilatih menggunakan data latih, yaitu Decision Tree, Random Forest, dan Self-Training Classifier. Model Self-Training Classifier menggunakan Decision Tree Classifier sebagai estimator.

Performa model diukur menggunakan akurasi (accuracy_score) dan classification report. Classification report menampilkan precision, recall, f1-score, dan support untuk setiap kelas.

Terakhir, data divisualisasikan menggunakan scatterplot dan histogram dengan Seaborn untuk beberapa fitur, yaitu mean radius, mean texture, mean perimeter, mean area, dan mean smoothness.

Hasil dari laporan teknis ini menunjukkan bahwa model Random Forest memberikan akurasi tertinggi di antara ketiga model dengan akurasi sebesar 0.953. Sedangkan, akurasi model Decision Tree dan Self-Training Classifier agak lebih rendah dengan akurasi berturut-turut sebesar 0.918 dan 0.930. Namun, perlu diingat bahwa kinerja model dapat bervariasi tergantung pada dataset yang digunakan dan parameter yang diatur.

3. Data Visualization

Visualisasi data merupakan langkah penting dalam memahami dan mengeksplorasi data. Dalam penelitian ini, kami menggunakan library seaborn untuk membuat plot berpasangan dari lima fitur rata-rata, yaitu jari-jari rata-rata, tekstur rata-rata, keliling rata-rata, luas rata-rata, dan rata-rata kehalusan. Plot berpasangan menunjukkan hubungan antara fitur yang berbeda dan distribusinya berdasarkan variabel target. Kami juga menggunakan scatterplot dan histogram untuk memvisualisasikan data dan variabel target.

4. Model Training and Evaluation

Data dibagi menjadi set pelatihan dan pengujian menggunakan fungsi train_test_split() dari pustaka sklearn.model_selection. Tiga pengklasifikasi, yaitu Pengklasifikasi Pohon Keputusan, Pengklasifikasi Hutan Acak, dan Pengklasifikasi Pelatihan Mandiri, dilatih di set pelatihan. Keakuratan pengklasifikasi kemudian dievaluasi menggunakan fungsi akurasi_score() dan laporan klasifikasinya dihasilkan menggunakan fungsi classification_report().

5. Results and Discussion

Pengklasifikasi pohon keputusan mencapai akurasi 0,9064, sedangkan pengklasifikasi hutan acak mencapai akurasi 0,9708. Pengklasifikasi pelatihan mandiri mencapai akurasi 0,9064. Dari hasil, terbukti bahwa pengklasifikasi hutan acak berkinerja lebih baik daripada pohon keputusan dan pengklasifikasi pelatihan mandiri.

6. Conclusion

Kesimpulannya, algoritma pembelajaran mesin dapat digunakan untuk mengklasifikasikan kanker payudara dan memprediksi hasilnya. Dataset Kanker Payudara digunakan untuk melatih dan mengevaluasi tiga pengklasifikasi yang berbeda, yaitu Pengklasifikasi Pohon Keputusan, Pengklasifikasi Hutan Acak, dan Pengklasifikasi Pelatihan Mandiri. Pengklasifikasi hutan acak mencapai akurasi tertinggi di antara ketiga pengklasifikasi. Hasilnya menunjukkan potensi algoritma pembelajaran mesin dalam diagnosis dan pengobatan kanker payudara.