TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

I Made Bayu Satria Wardhana 1103204145

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

1. Pendahuluan

Kanker payudara adalah salah satu jenis kanker yang paling sering didiagnosis pada wanita di seluruh dunia, dan deteksi dini memainkan peran penting dalam meningkatkan hasil pasien. Algoritma machine learning dapat membantu dalam mengklasifikasikan kanker payudara sebagai ganas atau jinak berdasarkan berbagai fitur yang diekstraksi dari gambar medis. Dalam proyek ini, kami menjelajahi penggunaan tiga algoritma klasifikasi yang berbeda: Decision Tree, Random Forest, dan Self-Training, untuk memprediksi hasil kanker payudara berdasarkan fitur tumor.

2. Data

Saya menggunakan dataset Breast Cancer Wisconsin (Diagnostic) dari library Scikit-Learn, yang berisi 569 contoh dan 30 fitur, termasuk nilai rata-rata, standar kesalahan, dan nilai maksimum dari atribut inti sel seperti radius, tekstur, perimeter, area, kehalusan, kepadatan, kecembungan, titik cekung, simetri, dan dimensi fraktal. Dataset juga termasuk variabel target yang menunjukkan apakah tumor ganas atau jinak.

3. Visualisasi Data

Kami menggunakan library Seaborn untuk membuat pair plot, yang menunjukkan hubungan berpasangan antara fitur yang berbeda dari dataset. Kami memplot fitur mean radius, mean texture, mean perimeter, dan mean area satu sama lain dan memberi warna titik berdasarkan variabel target. Plot menunjukkan bahwa fitur memiliki pemisahan yang jelas antara kedua kelas, sehingga cocok untuk klasifikasi.

4. Algoritma Klasifikasi

Ada tiga algoritma klasifikasi pada dataset: Decision Tree, Random Forest, dan Self-Training.

- Decision Tree

Decision tree adalah jenis algoritma pembelajaran terawasi yang digunakan untuk masalah klasifikasi. Mereka membangun model dalam bentuk struktur pohon di mana setiap simpul internal mewakili tes pada fitur, setiap cabang mewakili hasil tes, dan setiap simpul daun mewakili label kelas. Kami menggunakan kelas DecisionTreeClassifier dari library Scikit-Learn untuk melatih model pohon keputusan pada dataset kami. Kami mencapai akurasi 0,91 pada set pengujian.

- Random Forest

Random forest adalah jenis metode pembelajaran ensemble untuk klasifikasi, regresi, dan tugas lain yang beroperasi dengan membuat banyak pohon keputusan pada saat pelatihan dan mengeluarkan kelas yang merupakan mode dari kelas-kelas pohon individu. Kami menggunakan kelas RandomForestClassifier dari library Scikit-Learn untuk melatih model hutan acak pada dataset kami. Kami mencapai akurasi 0,96 pada set pengujian.

- Self-Training

Self-training adalah jenis metode pembelajaran semi-terawasi di mana algoritma memulai dengan dataset berlabel kecil dan secara iteratif menambahkannya dengan membuat prediksi pada data tidak berlabel dan menambahkan prediksi berkualitas tinggi ke dataset berlabel. Kami menggunakan kelas SelfTrainingClassifier dari library Scikit-Learn untuk melatih model self-training pada dataset kami. Kami menggunakan Decision Tree Classifier sebagai classifier dasar dan menetapkan jumlah iterasi maksimum menjadi 50. dengan akurasi mencapai sebesar 0,92 pada set pengujian.

5. Kesimpulan

Dalam proyek ini, kita menjelajahi penggunaan tiga algoritma klasifikasi: Decision Tree, Random Forest, dan Self-Training, untuk memprediksi hasil kanker payudara berdasarkan fitur dari tumor. Algoritma Random Forest memberikan kinerja terbaik dengan akurasi sebesar 0,96 pada set data uji. Algoritma Decision Tree dan Self-Training juga memberikan kinerja yang baik, dengan akurasi masing-masing sebesar 0,91 dan 0,92. Kitapun menggunakan library Seaborn untuk memvisualisasikan tren data, yang memberikan wawasan yang berguna tentang pemisahan dataset. Secara keseluruhan, proyek ini menunjukkan efektivitas algoritma pembelajaran mesin dalam memprediksi hasil kanker payudara berdasarkan fitur gambar medis.