TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

Rifqi Fadhila Shandi

1103202042

TK-44-04

PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS TEKNIK ELEKTRO
UNIVERSITAS TELKOM

2022/2023

A. Tujuan

Tujuan dari proyek ini adalah untuk mengeksplorasi dataset kanker payudara, memvisualisasikan trennya menggunakan Scikit Learn dan Seaborn, dan kemudian melakukan analisis data eksplorasi menggunakan klasifikasi Decision Tree, Random Forest, dan Self-Training.

B. Dataset

Dataset yang digunakan dalam proyek ini adalah dataset kanker payudara yang tersedia di perpustakaan Scikit Learn. Dataset ini berisi 569 sampel dengan 30 fitur masing-masing, dan variabel target adalah klasifikasi biner dari kanker payudara ganas atau jinak.

C. Persiapan Data

Pertama, akanmengimpor perpustakaan yang diperlukan seperti numpy, pandas, seaborn, matplotlib.pyplot, dan dataset kanker payudara dari Scikit Learn. Kami kemudian membagi dataset menjadi data berlabel dan tak berlabel dengan rasio 10:90 menggunakan seleksi acak. Data berlabel digunakan untuk melatih klasifikasi Decision Tree dan Random Forest, dan klasifikasi Self-Training menggunakan data berlabel dan tak berlabel.

D. Visualisasi Data

Selanjutnya, akan menggunakan Seaborn untuk memvisualisasikan tren dalam dataset kanker payudara. Disini akan menggunakan pair plot untuk memvisualisasikan hubungan berpasangan antara fitur, swarm plot untuk memvisualisasikan distribusi variabel target, dan box plot untuk memvisualisasikan distribusi fitur di seluruh variabel target.

E. Analisis Data Eksplorasi

Kemudian akan menggunakan tiga klasifikasi yang berbeda untuk melakukan analisis data eksplorasi pada dataset kanker payudara. Akan dilatih klasifikasi Decision Tree dan Random Forest pada data berlabel dan mengevaluasi kinerjanya pada data tak berlabel menggunakan laporan klasifikasi. Disini juga akan memvisualisasikan Decision Tree menggunakan plot_tree dan pentingnya fitur untuk Random Forest menggunakan plot batang.

Selanjutnya akan melakuka klasifikasi Self-Training untuk melakukan analisis data eksplorasi pada seluruh dataset. Klasifikasi Self-Training pertama-tama dilatih pada data berlabel dan kemudian memprediksi label kelas untuk data tak berlabel. Kemudian, klasifikasi menambahkan prediksi paling percaya ke data berlabel dan melatih kembali model. Proses ini berlanjut sampai seluruh data tak berlabel diberi label atau ambang kepercayaan tercapai. Terakhir akan mengevaluasi kinerja klasifikasi Self-Training pada data tak berlabel menggunakan laporan klasifikasi.

F. Hasil

Berikut adalah hasil dari program setelah dijalankan untuk masing – masing jenis klasifikasi adalah :

1. Decision Tree

   Metode klasifikasi Decision Tree digunakan untuk membangun model klasifikasi berdasarkan fitur-fitur pada data latih. Model ini kemudian diuji dengan data uji yang belum terlihat sebelumnya. Dalam eksperimen ini, Decision Tree menghasilkan nilai presisi sebesar 0,91 pada data uji.

2. Random Forest

   Metode klasifikasi Random Forest adalah teknik ensemble learning yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi klasifikasi. Pada eksperimen ini, Random Forest menghasilkan nilai presisi sebesar 0,97 pada data uji.Untuk memahami lebih lanjut tentang fitur mana yang lebih signifikan dalam pengambilan keputusan model, feature importance dibuat menggunakan seaborn dan matplotlib.

3. Self-Training

   Metode klasifikasi Self-Training adalah teknik semi-supervised learning yang menggunakan data latih dan data uji yang tidak bertanda untuk memperbaiki akurasi klasifikasi. Dalam eksperimen ini, metode Self-Training menghasilkan nilai presisi sebesar 0,94 pada data uji.

G. Kesimpulan

Dari tiga metode klasifikasi yang digunakan, Random Forest menghasilkan nilai presisi tertinggi pada data uji. Namun, Self-Training juga menunjukkan kinerja yang baik dalam mengklasifikasikan sampel pada data uji. Visualisasi data tren dan feature importance juga membantu dalam pemahaman fitur-fitur yang lebih penting dalam membuat keputusan klasifikasi.