TECHNICAL REPORT MACHINE LEARNING

"Breast Cancer Dataset Exploration Using Python and Sckit-Learn"

Disusun oleh :

Arfiq Rimeldo

1103202102

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

UNIVERSITAS TELKOM

2023

Pendahuluan

Kanker payudara adalah suatu kondisi di mana sel-sel abnormal pada jaringan payudara tumbuh secara tidak terkendali dan dapat menyebar ke bagian tubuh lainnya.Pada technical report ini kita akan melakukan penelitian terhadap dataset kanker payudara menggunakanpython dan sckit-learn.

Deskripsi Dataset

Dataset ini berisi informasi tentang sampel tumor payudara dari pasien yang menderita kanker payudara serta informaasi mengenai ukuran dan bentuk tumornyaDataset ini terdiri dari 569 sampel dengan 30 fitur,salah satunya fiturnya yaitu “target” merupakan variabel biner yang menunjukkan keganasan tumor.

Visualisasi Data

Visualisasi data adalah suatu teknik untuk mempresentasikan data dan informasi dalam bentuk grafik, diagram, atau gambar yang dapat dengan mudah dipahami dan diinterpretasikan.Disini saya menggunakan teknik scatterplot dan heatmap,Scatterplot digunakan untuk menjelaskan hubungan anatara 2 variabel,yaitu mean radius dan mean area serta menggunakan target untuk melhat tingkat keganansan dari tumor pada mean area dan mean radius tersebut.

Heatmap digunakan untuk memvisualisasikan data numerik dalam bentuk matriks yang diberi warna,sehingga memudahkan pengamatan dan analisis data.Setiap sel dalam matriks merepresentasikan nilai numerik, dan warna yang diberikan pada setiap sel diatur berdasarkan skala warna tertentu

Decision Tree

Decision tree adalah model pembelajaran mesin yang menghasilkan pohon keputusan dari data yang diberikan.Setiap kotak dalam pohon merepresentasikan suatu keputusan atau prediksi,dan cabang-cabang dari kotak tersebut merepresentasikan kemungkinan nilai dari fitur-fitur yang digunakan untuk membuat keputusan tersebut.Decision tree disini mencapai tingkat akuarsi 94%,presisi 94%,recall 94% serta f1-score juga 94%.

Random Forest

Random forest adalah algoritma pembelajaran mesin yang memanfaatkan teknik ensemble learning,yaitu menggabungkan beberapa model pembelajaran mesin yang berbeda untuk meningkatkan performa prediksi.Random forest terdiri dari beberapa pohon keputusan yang dibangun secara acak,dimana setiap pohon dihasilkan dari subset data dan subset fitur yang dipilih secara acak.Random forest disini memiliki tingkat akurasi 96%,presisi 94% dan recal 94%.

Self Training

Self-training adalah teknik pada machine learning,dimana model pembelajaran mesin dilatih dengan menggunakan data yang terdiri dari label dan tanpa label.Fungsi self-training adalah untuk meningkatkan performa prediksi model pada data yang tidak memiliki label yang cukup.Self training disini memiliki tingkat akurasi 97%.

Kesimpulan

Dari technical report ini dapat disimpulkan suatu tumor yang menyebabkan kanker payudara itu merupakan tumor ganas atau tumor jinak dengan menggunakan berbagai macam algoritma dengan tingkat akurasi yang berbeda beda namun bisa dibilang semuanya memiliki tingkat akurasi yang tinggi.