UJIAN TENGAH SEMESTER MACHINE LEARNING

"Analisis Dataset Kanker Payudara menggunakan Python dan Scikit-Learn Framework"

Disusun oleh:

Fariz Rahman Ramadhan
1103204046

PROGRAM STUDI TEKNIK KOMPUTER
FAKULTAS TEKNIK ELEKTRO
UNIVERSITAS TELKOM
2023

I. PENDAHULUAN

Kanker payudara adalah kanker yang terbentuk di jaringan payudara. Kanker payudara terjadi ketika sel-sel pada jaringan di payudara tumbuh secara tidak terkendali dan mengambil alih jaringan payudara yang sehat dan sekitarnya. Pada kesempatan kali ini kita mempunyai Dataset Kanker Payudara yang akan digunakan untuk dianalisis menggunakan Python, Scikit-Learn, dan juga Seaborn. Untuk jenis-jenis Algoritma yang digunakan diantaranya yaitu: Decision Tree, Random Forest, dan juga Self Training.

II. DATASET

Pada Analisis kali ini dataset yang digunakan adalah Dataset Kanket Payudara dari Breast Cancer Winconsin Dataset, yang memiliki isi tentang informasi sampel kanker payudara dari pasien yang menderita kanker payudara. Dataset ini berisikan 569 sampel dengan 30 fitur yang menjelaskan masing-masing karakteristik sel.

III. VISUALISASI DATA

Visualisasi data adalah suatu cara untuk menyajikan data dalam bentuk grafis, gambar, atau diagram. Data yang tersedia divisualisasi menggunakan Teknik Scatterplot dan juga Pairplot. Scatterplot adalah suatu jenis grafik yang digunakan untuk menampilkan korelasi antara dua variabel numerik pada sumbu-x dan sumbu-y. Pada Scatterplot saya gunakan data ‘mean perimeter’ dan juga ‘mean texture’ sehingga didapatkan korelasi antara kedua variable tersebut.

Selain itu Pairplot juga digunakan untuk menunjukkan hubungan antara beberapa pasang variabel dalam dataset yang sama. Variabel-variabel yan digunakan seperti mean symmetry, mean concavity, mean compactness, dan juga mean smoothness.

IV. ANALISIS DATA

Pada analisis data digunakan Decision Tree, Random Forest, dan juga Self Training. Decision tree (pohon keputusan) adalah algoritma pembelajaran mesin yang menghasilkan model prediktif dalam bentuk pohon yang terdiri

dari serangkaian simpul dan cabang. Pada setiap simpul, terdapat suatu pertanyaan atau kondisi yang digunakan untuk membagi data ke dalam dua atau lebih kelompok yang homogen berdasarkan nilai dari suatu fitur atau variabel tertentu. Decision Tree menghasilkan akurasi sebesar 92%, presicion 95%, recall 92%, dan F1-score 94%.

Random Forest (hutan acak) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk klasifikasi, regresi, dan pengelompokan data. Random Forest dapat meningkatkan akurasi prediksi dan mengurangi overfitting. Hasil dari Random Forest yaitu akurasi 98%, presisi 99%, recall 98%, f1-score 99%.

Self-training atau self-learning adalah metode pembelajaran mesin semi-terawasi (semi-supervised) di mana model belajar dari data yang sebagian besar tidak berlabel (unsupervised) dan sedikit data berlabel (supervised), self-training pada analisis kali ini memiliki tingkat keakurasian sekitar 98%.

V. KESIMPULAN

Pada Analisis kali ini yang menggunakan Dataset Kanket Payudara dari Breast Cancer Winconsin Dataset, Setelah dilakukan Analisis Data menggunakan tiga metode yang berbeda yaitu Decision Tree, Random Forest, dan juga Self-Train, metode Random Forest memiliki tingkat keakurasian paling tinggi yaitu sekitar 98%-99%. analisis ini menunjukkan bahwa algoritma pembelajaran mesin dapat digunakan untuk memprediksi kanker payudara dan memberikan hasil yang akurat. Selain itu, visualisasi data juga sangat penting dalam analisis data untuk membantu memahami korelasi antara variabel-variabel yang berbeda. Dataset kanker payudara dari Breast Cancer Wisconsin dapat menjadi sumber informasi yang berguna bagi penelitian lebih lanjut tentang kanker payudara dan pengembangan algoritma pembelajaran mesin untuk memprediksi kanker payudara.