Technic Report

Visualize and Explore Data Using Breast Cancer Dataset

Tugas Untuk Memenuhi Mata Kuliah *Machine Learning*

Disusun Oleh:

Nama : Rai Barokah Utari

NIM : 1103200066

PROGRAM STUDI SI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

TELKOM UNIVERSITY

BANDUNG

2023

I. Pendahuluan

Kanker payudara adalah salah satu jenis kanker yang paling umum terjadi pada wanita di seluruh dunia. Kanker ini terjadi ketika sel-sel di dalam payudara tumbuh secara abnormal dan tidak terkontrol. Kanker payudara dapat terjadi pada pria, meskipun kasusnya jauh lebih jarang. Kanker payudara adalah suatu jenis tumor ganas yang berkembang pada sel-sel payudara. Kanker ini dapat tumbuh jika terjadi pertumbuhan yang abnormal dari sel-sel pada payudara. Sel-sel tersebut membelah diri lebih cepat dari sel normal, yang kemudian membentuk benjolan atau massa. Pada stadium yang lebih parah, sel-sel abnormal ini dapat menyebar melalui kelenjar getah bening ke organ tubuh lainnya. Kanker payudara dapat menyebabkan berbagai gejala seperti benjolan pada payudara, nyeri pada payudara, perubahan bentuk atau ukuran payudara, perubahan pada kulit payudara, dan perubahan pada puting susu. Meskipun demikian, tidak semua benjolan pada payudara adalah kanker payudara. Sebagian besar benjolan pada payudara adalah jinak atau non-kanker.

Pemeriksaan dini dan deteksi kanker payudara sejak dini dapat memperbaiki prognosis pasien dan memberikan peluang kesembuhan yang lebih baik. Pemeriksaan dapat dilakukan melalui pemeriksaan fisik, mamografi, USG, dan MRI. Selain itu, penelitian terus dilakukan untuk mengembangkan metode deteksi dan pengobatan kanker payudara yang lebih efektif dan terjangkau. Pada *technical report* ini, akan dilakukan beberapa visualisasi data dan eksplorasi menggunakan beberapa teknik yang ada di dalam machine learning.

II. Pengumpulan Data

Tahap pertama yang dilakukan yaitu Import Library dan Load Dataset Pada bagian ini, dilakukan import beberapa library yang dibutuhkan untuk pengolahan data seperti numpy, pandas, matplotlib, seaborn, dan beberapa library dari sklearn, kemudian Eksplorasi Data Pada bagian ini, dilakukan beberapa eksplorasi data seperti melihat informasi umum dari data seperti jumlah baris dan kolom, tipe data dari masing-masing kolom, serta statistik deskriptif dari masing-masing kolom, Pemisahan Data pada bagian ini, dilakukan pemisahan data menjadi data train dan data test menggunakan train_test_split. Dan terakhir, kita dapat menggunakan heatmap untuk melihat tingkat korelasi antara setiap features yang ada di dalam dataset tersebut.

III. Eksplorasi Data

Eksplorasi data dengan menggunakan teknik Random Forest untuk menggambarkan feature importances dari random forest classifier pada kanker payudara ini yang selanjutnya dapat kita lakukan visualisasi pada setiap features yang ada. Selanjutnya, kita juga dapat menentukan hasil dari akurasi skor, classification report, dan confusion matrix untuk model predictionnya yang selanjutnya dapat kita tampilkan dan mendapatkan hasilnya. Eksplorasi data terakhir, adalah dengan menggunakan self-training. Hal yang pertama kali kita lakukan adalah dengan melakukan split data untuk dipisah menjadi training sets dan testing sets. Selanjutnya, disini kita dapat mengubah data cancer tersebut menjadi Dataframe dan dapat kita lakukan plot untuk melihat distribusi dari target yang ada. teknik self-training classifier untuk mengklasifikasikan dataset breast cancer tersebut.

Teknik Self-Training yang dilakukan adalah dengan menggunakan semi-supervised learning di mana model akan diawali dengan beberapa data yang sudah dilabeli dan kemudian menggunakan data yang tidak dilabeli untuk melatih dirinya sendiri secara berulang sampai konvergen. Tujuan dari eksperimen ini adalah untuk membandingkan performa klasifikasi dan jumlah sampel yang dilabeli dengan mengatur parameter threshold pada Self-Training Classifier. Eksperimen akan ini menghasilkan dua grafik. Grafik pertama menunjukkan nilai akurasi rata-rata dari 3-fold cross validation dengan menggunakan threshold yang berbeda-beda. Grafik ketiga menunjukkan jumlah iterasi yang diperlukan oleh model untuk konvergen dengan menggunakan threshold yang berbeda-beda.

IV. Kesimpulan

Dapat disimpulkan bahwa dataset Breast Cancer Wisconsin (Diagnostic) dapat digunakan untuk memprediksi jenis tumor dengan akurasi yang cukup tinggi menggunakan model KNN dengan memilih fitur-fitur yang tepat dan melakukan transformasi data yang diperlukan. Selain itu, pada eksplorasi data juga ditemukan beberapa hal yang perlu diperhatikan dalam pemrosesan data, seperti normalisasi data dan deteksi outliers.