

TECHNICAL REPORT

**VISUALIZE AND EXPLORE DATA USING BREAST CANCER
DATASET**



TUGAS UNTUK MEMENUHI MATA KULIAH

MACHINE LEARNING

Oleh:

Muhammad Rizky Pradhitia

1103204192

PROGRAM STUDI SI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO

TELKOM UNIVERSITY

BANDUNG

2023

A. Pendahuluan

Kanker Payudara merupakan tumor ganas yang berasal dari sel-sel payudara dan penderita yang mengalami penyakit ini mayoritas adalah Wanita. Deteksi dan diagnosis yang akurat dari permasalahan kanker payudara sangat penting untuk pengobatan yang efektif. Dalam hal ini, algoritma pembelajaran mesin telah banyak digunakan dalam penelitian medis untuk mengklasifikasikan kanker payudara berdasarkan berbagai fitur yang diekstraksi dari data medis. Pada technical report ini, akan dilakukan eksplorasi dan visualisasi dataset dari kanker payudara. Dataset yang digunakan berisi berbagai fitur yang telah diekstraksi dari pasien yang memiliki penyakit kanker payudara, seperti rata-rata radius, rata-rata tekstur, rata-rata keliling, rata-rata luas, dan rata-rata kehalusan, beserta label target yang menunjukkan apakah tumor tersebut ganas atau jinak.

B. Visualisasi Data

Pertama saya mengambil dataset breast cancer Wisconsin dengan menggunakan metode `load_breast_cancer` yang diambil dari library `sklearn` dan diatur untuk diambil dalam format DataFrame menggunakan `as_frame = True`. hasilnya akan disimpan pada variabel data. Untuk memvisualisasikan dataset tersebut saya menggunakan beberapa fungsi seperti `countplot`, `pairplot`, dan `heatmap` yang disediakan dari library `seaborn`.

Untuk, `countplot` digunakan untuk membuat histogram target variabel. Dalam histogram ini akan mengindikasikan apakah tumor jinak atau ganas. Untuk `pairplot` digunakan untuk membuat plot pasangan antara beberapa fitur yang digunakan, yaitu mean radius, mean texture, mean perimeter, mean area, dan mean smoothness. Plot ini memungkinkan untuk melihat bagaimana fitur-fitur ini berkorelasi satu sama lain dan juga dengan variabel target, dan terakhir fungsi yang saya gunakan yaitu `heatmap`. `Heatmap` digunakan untuk membuat sebuah matriks korelasi. Dalam kasus ini, matriks korelasi menunjukkan seberapa kuat korelasi antara setiap pasang fitur dalam dataset. Dalam penggunaan fungsi `countplot`, dapat dilihat bahwa dataset seimbang dengan proporsi kasus ganas dan jinak.

C. Eksplorasi Data

Untuk Eksplorasi data, model pembelajaran mesin yang saya gunakan pertama adalah model decision tree. Model ini digunakan untuk melakukan klasifikasi pada dataset kanker payudara. Pertama saya membagi dataset menjadi data latih dan data uji dengan perbandingan 80:20 menggunakan `train test split` dari library `scikit-learn`. Kemudian saya membuat objek sebagai klasifikasi decision tree dan kemudian melakukan pelatihan model menggunakan data latih `X_train` dan label `y_train`. Setelah dilakukan proses

klasifikasi decision tree, saya melakukan proses pruning (memangkas pohon keputusan) untuk menghindari *overshifting*. Untuk mendapatkan Jalur kompleksitas biaya yang optimal dilakukan dengan menggunakan **cost complexity pruning path**. Kemudian nilai alpha disimpan pada `ccp_alphas`. setelah itu dilakukan iterasi sebanyak `ccp_alpha` dan untuk setiap `ccp_alpha` dilakukan pembuatan decision tree baru dengan nilai `ccp_alpha`. decision tree yang baru kemudian di fit ke training set dan dimasukkan ke dalam list.

Selanjutnya, dilakukan perhitungan train score dan test score untuk setiap decision tree pada list dan nilai nya akan disimpan dalam list `train_scores` dan `test_scores`. Hasil dari perhitungan `train_score` dan `test_score` akan digunakan untuk membuat plot yang menampilkan grafik akurasi terhadap nilai alpha. Grafik ini dapat digunakan untuk menentukan nilai `ccp_alpha` yang optimal untuk digunakan pada decision tree classifier. Setelah mendapatkan nilai `ccp_alpha` yang optimal, kemudian membuat decision tree baru dengan nilai `ccp_alpha` yang optimal dan dimasukkan ke dalam variabel `clf_pruned`. Decision tree baru kemudian di fit ke training set dan diuji akurasinya pada testing set. Hasil akurasi yang saya dapatkan sebesar 95%.

Untuk model yang kedua saya menggunakan random forest untuk melakukan klasifikasi pada data dan memprediksi label target berdasarkan pada data. Pertama, dilakukan pembuatan model random forest. Selanjutnya dilakukan perhitungan feature importances dari model random forest. Nilai nilai tersebut kemudian diurutkan secara descending. setelah itu, feature names yang telah diurutkan disesuaikan sehingga sama dengan feature importances yang telah diurutkan. Selanjutnya menghitung permutation importance. Metode ini bekerja dengan secara acak mengacak nilai setiap feature dengan menggunakan seberapa besar penurunan akurasi model sebagai hasilnya. Setelah menghitung feature, model memprediksi data label menggunakan metode predict dari pengklasifikasian random forest. Kemudian, model menghitung akurasi score dengan hasil 96%. Untuk confusion matrix menunjukkan jumlah true positif, true negative, false positif, dan false negative.

Untuk model yang terakhir saya menggunakan self training. Model dasar yang digunakan adalah **SVC**. yang pertama dilakukan adalah memuat dataset cancer payudara dengan mengacaknya dengan random state. Untuk menguji self-training, beberapa sampel diubah menjadi -1, yang menandakan sampel tidak ditandai. selanjutnya, menyiapkan variabel untuk menyimpan hasil percobaan self-training classifier pada dataset. Kode membagi dataset ke dalam tiga bagian untuk pengujian dan iterasi dibuat melalui threshold. Kemudian, variabel hasil digunakan untuk menyimpan hasil akurasi, jumlah sampel, dan iterasi pada setiap nilai threshold. Di dalam loop, self-training classifier dilatih pada subset yang ditandai, dan kemudian digunakan untuk memprediksi label

sampel yang tidak ditandai pada subset lain. Kemudian, akurasi dan jumlah sampel yang diberi label dan iterasi yang dihitung dan disimpan dalam variabel hasil. Kemudian akan memplot hasil dari akurasi model pada setiap nilai threshold, menunjukkan jumlah sampel yang diberi label pada setiap nilai threshold, dan menunjukkan jumlah iterasi pada setiap nilai threshold. Hasil dari plot ini akan membriakan informasi tentang bagaimana peningkatan jumlah sampel yang diberi label mempengaruhi akurasi dan iterasi dalam self-training classifier

D. Kesimpulan

Dalam laporan teknis ini, Kode melakukan analisis data eksplorasi, implementasi algoritma, dan evaluasi algoritma. Algoritma yang diimplementasikan termasuk decision tree classifier, pruned decision tree classifier, random forest classifier, dan self-training classifier. dan telah membangun dan membandingkan beberapa model pembelajaran mesin untuk klasifikasi kanker payudara berdasarkan fitur-fitur dari data medis. Hasil eksperimen menunjukkan bahwa hutan random, SVM, dan KNN memiliki akurasi yang lebih tinggi dibandingkan dengan pohon keputusan yang telah dipangkas. Namun, pemilihan model terbaik masih bergantung pada karakteristik dataset dan kebutuhan aplikasi yang spesifik. Penggunaan algoritma pembelajaran mesin dalam diagnosis kanker payudara dapat membantu dalam deteksi dini dan pengobatan yang efektif, namun perlu validasi